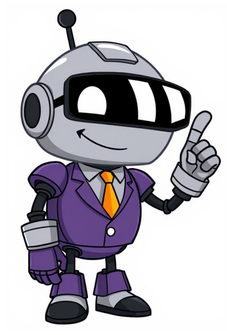
Continue



Presto 即 Presto 数据库 (PrestoDB)。它是一种开源的分布式 SQL 查询引擎，可以查询来自不同数据源的大数据集，以便企业能够大规模地解决数据问题。 它为各种规模的组织提供一种快速、有效的方法来分析来自各种来源（包括本地部署系统和云）的大数据。Presto 还可以帮助企业使用当前的 SQL 功能查询 PB 级数据，而无需学习新语言。 如今，Presto 最常用于在 Hadoop 等常见数据存储平台上运行查询，支持用户管理多种查询语言以及数据库和存储的接口。 在数字时代，大数据分析正迅速成为各规模各行业企业的核心能力。收集、存储和分析与业务流程、客户偏好和市场趋势相关的海量数据的能力是非常有价值的。Presto 对数据分析的最大重要性在于不受数据存储位置的制约，无需先将数据移至更结构化的系统（例如数据仓库或数据湖）。 在每周的 Think 时事通讯中，发现专家精选的有关 AI、云等的洞察分析和新闻。 Presto 已成为数据科学家和工程师处理多种查询语言、孤岛式数据库和不同类型存储的常用工具。它的高性能功能使用户能够使用简单的 ANSI SQL 界面实时查询大量数据，无论数据位于何处。Presto 在执行海量数据查询方面的速度和性能使其成为众多全球规模最大的公司不可或缺的工具，包括 Facebook、Airbnb、Netflix、Microsoft、Apple (iOS) 和 AWS（Athena 和 Amazon s3）。 Presto 架构的独特之处在于，无论数据存储在何处，Presto 都可以查询数据，比其他类似解决方案更具可扩展性和效率。Presto 查询支持工程师使用数据，而无需将数据从一个位置以物理方式移动到另一个位置。当组织处理需要存储和分析的数据量不断增加时，这是一项重要功能。 Presto 旨在使数据科学家和工程师能够交互式查询大量数据，且无需考虑存储来源或类型。由于 Presto 不存储数据，而是与单独的数据库进行查询通信，因此它比竞争对手更灵活，并且可以根据组织不断变化的需求快速扩展或缩减查询。根据 IBM 白皮书，Presto 已针对商业智能 (BI) 工作负载进行优化，它可以帮助企业优化数据仓库的定价，并将成本降低高达 50%。 以下是使用 Presto 工作流程的一些主要优点： 更低的成本：随着数据仓库规模的扩大和查询用户数量的增加，企业成本迅速增加的情况并不少见。然而，Presto 针对大量小型查询进行了优化，可以轻松查询任意数量的数据，同时降低成本。此外，由于 Presto 是开源的，因此部署它不需要任何费用，这可以为希望处理大量数据的企业节省大量成本。 提高可扩展性：工程师通常在单个数据湖存储系统上设置多个引擎和语言，因此将来有必要重新构建平台并限制解决方案的可扩展性。借助 Presto，所有查询均使用通用 ANSI SQL 语言和界面进行，因此无需重新构建平台。此外，Presto 可用于处理少量数据和大量数据，并可轻松从一两个用户扩展到数千个用户。Presto 使用独特的 SQL 方言和 API 部署多个计算引擎，因此成为扩展工作负载的理想工具，这些工作负载可能过于复杂且耗时，超出工程师和数据科学家团队的处理能力。 更好的性能：许多在 Hadoop 上运行 SQL 的查询引擎计算性能都会受到限制，因为它们被构建为将结果写入磁盘，但 Presto 的分布式内存模型使其能够针对大型数据集同时运行大量交互式查询。按照经典的大规模并行处理 (MPP) 设计，Presto 会在单个工作节点上调度尽可能多的查询，并使用内存中流式洗牌来进一步提高处理速度。在内存中执行任务使得各阶段之间的磁盘写入和读取变得多余，缩短了每次查询的执行时间，使 Presto 成为比竞争对手延迟更低的选择。 更好的灵活性：Presto 对其所有数据源（包括 Cassandra、Kafka、MySQL、Hadoop 分布式文件系统 (HDFS)、PostgreSQL 等）都采用了即插即用模式，与缺乏这种功能的其他同类工具相比，跨数据源查询更快、更简单。此外，Presto 灵活的架构意味着它并不局限于一家供应商，而是可以在大多数 Hadoop 发行版上运行，这使 Presto 成为可移植性最强的工具之一。 虽然 Presto 并不是开发人员和数据工程师唯一的 SQL-on-Hadoop 选项，但其独特的架构将查询功能与数据存储分开，使 Presto 成为最灵活的选项之一。与其他工具不同，Presto 将查询引擎与数据存储分开，并使用连接器在它们之间进行通信。与其他工具相比，这一附加功能为工程师在如何使用 Presto 构建解决方案方面提供了更大的灵活性。 Presto 使用 MPP 数据库管理系统，其中一个协调器节点与其他节点协同工作。Presto 生态系统由三种服务器类型组成：协调器服务器、工作服务器和资源管理器服务器。 协调器：协调器被视为 Presto 的"大脑"。它负责一些最关键的任务，包括解析语句、规划查询和管理 Presto 工作器节点；最终负责从工作器节点检索数据并将结果传递给客户端。 工作程序：工作程序负责从工作程序节点收集数据，并保证自身与连接器之间顺利交换数据。 资源管理器：资源管理器从所有协调器和工作程序节点收集数据并创建全局视图（或称为"Presto 集群"）。 Presto 协调器 SQL 服务器收到用户的 SQL 查询后，协调器首先使用自定义查询来解析、规划和调度覆盖其他节点的分布式计划。Presto REST API 提交查询语句供服务器执行，并为客户端检索结果。Presto 支持标准 ANSI SQL 语义，包括连接、查询、子查询和聚合。 Presto 编译完查询后，会将请求解析到工作节点之间的不同阶段。 由于 Presto 基于数据抽象概念而构建，因此它可以扩展到任何数据源，并且可以轻松查询数据湖、数据仓库和关系数据库等数据源。数据抽象是一种编程过程，通过将数据的表示与其物理存储分开，可以更有效地存储和操作数据。这种抽象允许查询引擎专门关注与查询相关的数据方面。使用数据抽象过程，无论数据存储在何处，都可以进行查询，而不是在将数据迁移到另一个分析系统后进行查询。 Presto 最初是在 Facebook 开发的，用于在大型 Apache Hadoop 数据仓库上运行交互式查询，Presto 的开发人员始终将其设想为开源软件，并寻求将 Presto 免费用于商业用途，以便任何人都可以使用它进行数据分析和数据管理。2013 年，Presto 在 GitHub 上被设置为开源软件，拥有 Apache Software 许可证的任何人都可以下载。2019 年，Presto 开发团队的三名原始成员离开了该项目，并创立了 Presto 的一个"分支"，称为 Presto Software Foundation（或更多称为 prestosql）。 Linux 基金会和其他开源社区为希望获得认证的工程师和开发人员提供英语等语言的 Presto 网络研讨会和培训。这些论坛也是了解 Presto 最新资讯的不错平台。 Presto 使组织能够快速高效地查询大型数据存储库和 NoSQL 数据库，以实现各种业务目的。以下是一些最常见的用例： 临时查询 Presto 能够为各种业务目的提供快速的数据探索和简单的报告。使用热门的 Presto 连接器，例如 Hive、MongoDB 或 Cassandra，用户可以查询他们感兴趣的数据并在数秒内获得结果。凭借其速度和灵活性，无论数据集位于何处，Presto 支持用户迭代并进一步探索。 以下是 Presto 可以连接的一些使用最广泛的数据存储库： BigQueryHDFS云存储Cloud SQL for MySQLApache Cassandra 或 Kafka 云和混合云部署 根据 RedHat 2021 年的性能评估，企业越来越多地使用混合云环境，这给云原生存储带来了越来越大的压力，而"当今最快的分布式查询引擎"Presto 正是这种压力下的理想选择。1将工作负载从本地部署环境转移到云或混合云基础设施有很多好处，包括提高性能和可扩展性。Presto 的体系架构使其成为此类部署的有力选择，因为它可以在几分钟内启动，无需额外的预配、配置或调整。 机器学习 (ML) Presto 帮助工程师以高效的方式准备数据并执行特征工程和提取，确保其为机器学习 (ML) 做好准备。它拥有大量的连接器、SQL 引擎和查询功能，是工程师快速、轻松访问大量数据的理想选择。此外，Presto 还有专门为聚合等 ML 功能设计的工具，允许数据科学家训练支持向量机 (SVM) 分类器和回归器，以解决监督学习问题。 报告 Presto 允许从多个源查询数据，生成用于 BI 目的的单个、易于访问的报告或仪表板。Presto 简单易用，分析师无需工程师的帮助即可进行查询并创建报告。 分析 Presto 使分析师能够直接在数据湖上对结构化和非结构化数据进行查询，而无需执行数据转换过程。 数据准备 收集和准备数据的过程可能成本高昂且效率低下。数据科学家每天需要花费数小时来收集和准备数据，然后才能进行分析。Presto 能够快速准确地自动执行此过程，因此数据科学家和工程师可以将更多时间集中在更有价值的任务上。 Presto, or Presto database (PrestoDB), is an open-source, distributed SQL query engine that can query large data sets from different data sources, enabling enterprises to address data problems at scale. Presto gives organizations of all sizes a fast, efficient way to analyze big data from various sources including on-premise systems and the cloud. It also helps businesses query petabytes of data using their current SQL capabilities, without having to learn a new language. Today, Presto is most commonly used for running queries on Hadoop and other common data storage providers, enabling users to manage multiple query languages and interfaces to databases and storage. In the digital age, big data analytics is fast becoming a core competency for enterprises regardless of size or industry. The ability to gather, store and analyze large amounts of data relating to business processes, customer preferences and market trends is extremely valuable. Presto's primary importance to data analytics is its ability to analyze data regardless of where the data is being stored, and without having to move it into a more structured system first, such as a data warehouse or data lake. Discover expertly curated insights and news on AI, cloud and more in the weekly Think Newsletter. Presto has become a popular tool for data scientists and engineers dealing with multiple query languages, siloed databases and different types of storage. Its high-performance capabilities enable users to query large volumes of data in real-time, regardless of where the data is located, using a simple ANSI SQL interface. Presto's speed and performance at executing queries on large volumes of data have made it an indispensable tool for some of the largest companies in the world including Facebook, Airbnb, Netflix, Microsoft, Apple (iOS) and AWS (Athena and Amazon s3). Presto architecture is unique in that it is built to query data no matter where the data is being stored, making it more scalable and efficient than other, similar solutions. Presto queries allow engineers to use data without having to physically move it from location to location. This is an important capability to have as organizations deal with an ever-increasing amount of data they need to store and analyze. Presto was built to empower data scientists and engineers to interactively query vast amounts of data regardless of the source or type of storage. Because Presto doesn't store data, but rather communicates with a separate database for its queries, it is more flexible than its competitors and can scale queries up or down swiftly based on the shifting needs of the organization. According to an IBM whitepaper, Presto, optimized for business intelligence (BI) workloads, can help enterprises optimize the pricing of their data warehouses and reduce costs by up to 50 percent. Here are some of the key benefits to using a Presto workflow: Lower costs: As the size of data warehouses and the number of users conducting queries grows, it's not uncommon for enterprises to see their costs rapidly increase. Presto, however is optimized for large amounts of small queries, making it easy to query any amount of data while also keeping costs down. Also, since Presto is open-source, there are no fees associated with deploying it, which can result in significant savings for enterprises looking to process large volumes of data. Increased scalability: It's common for engineers to set up multiple engines and languages on a single data lake storage system, which can make it necessary to re-platform in the future and limit the scalability of their solution. With Presto, all queries are conducted using the universal ANSI SQL language and interface, making re-platforming redundant. Additionally, Presto can be used for both small and large amounts of data and easily scaled up from one or two users to thousands. Presto deploys multiple compute engines with unique SQL dialects and APIs, making it an ideal tool for scaling workloads that could be too complex and time-intensive for teams of engineers and data scientists to handle. Better performance: While many query engines that run SQL on Hadoop are restricted in their compute performance because they are built to write their results to disk, Presto's distributed in-memory model enables it to run large amounts of interactive queries at once against large data sets. Following a classic massively parallel processing (MPP) design, Presto schedules as many queries as it can on a single worker node and uses in-memory streaming shuffle to increase its processing speeds even more. Executing tasks in-memory makes writing and reading from the disk between stages redundant and shortens the time of each query execution, making Presto a lower latency option than its competitors. Improved flexibility: Presto uses a plug-and-play model for all its data sources including Cassandra, Kafka, MySQL, Hadoop distributed file system (HDFS), PostgreSQL and others, making querying across them faster and easier than with other comparable tools that lack this functionality. Also, Presto's flexible architecture means it isn't restricted to a single vendor but runs on most Hadoop distributions, making it one of the most portable tools available. While Presto isn't the only SQL-on-Hadoop option available to developers and data engineers, its unique architecture that keeps query functionality separate from data storage makes it one of the most flexible. Unlike other tools, Presto siloes off the query engine from the data storage and uses connectors to communicate between them. This added functionality gives engineers more flexibility than other tools in how they construct solutions using Presto. Presto uses an MPP database management system with one coordinator node that works in tandem with other nodes. A Presto ecosystem is made up of three server types, a coordinator server, worker server and resource manager server. Coordinator: A coordinator is considered the "brains" of a Presto installation. It is responsible for some of the most critical tasks including parsing statements, planning queries and managing Presto worker nodes. Ultimately, it is responsible for retrieving data from the worker nodes and delivering the results to the client. Worker: The worker is responsible for gathering data from the worker nodes and ensuring the smooth exchange of data between itself and the connectors. Resource manager: The resource manager gathers data from all the coordinator and worker nodes and creates a global view, or a "Presto cluster". When the Presto coordinator SQL server receives an SQL query from a user, the first thing it does is use a custom query to parse, plan and schedule a distributed plan across the other nodes. Presto REST API is used to submit query statements for execution on a server and to retrieve the results for the client. Presto supports standard ANSI SQL meanings, including joins, queries, sub-queries and aggregations. Once it has compiled the query, Presto parses the request into different stages between the worker nodes. Since Presto was built on the concept of data abstraction, it is extensible to any data source and can easily query data sources such as data lakes, data warehouses and relational databases. Data abstraction is a programming process that allows data to be stored and manipulated more efficiently by separating its representation from its physical storage. This abstraction allows for the query engine to focus exclusively on the aspects of the data that are relevant to its query. Using the process of data abstraction, data is queried wherever it is being stored rather than once it has been moved into another analytics system. Initially developed at Facebook to run interactive queries on a massive Apache Hadoop data warehouse, Presto's developers always envisioned it as open-source software and sought to make it free for commercial use so anyone could use it for data analytics and data management. In 2013, it was open-sourced on GitHub for anyone to download under the Apache Software license. In 2019, three of the original members of the Presto development team left the project and founded a "fork" of Presto known as the Presto Software Foundation, or more commonly, as prestosql. The Linux foundation and other open-source communities offer webinars and training on Presto in English and other languages for engineers and developers looking to obtain certification. These forums are also a good place to visit to find out what's new in Presto. Presto enables organizations to query large-scale data repositories and NoSQL databases quickly and efficiently for a variety of business purposes. Here are some of its most common use cases: Ad-hoc queries Presto enables speedy data exploration and straightforward reporting for a variety of business purposes. Using popular Presto connectors, such as Hive, MongoDB or Cassandra, users can query data they have an interest in and get results in seconds. With its speed and flexibility, Presto empowers users to iterate on and further explore data sets, regardless of where they reside. Here are some of the most widely used data repositories that Presto can connect to: BigQueryHDFSCloud StorageCloud SQL for MySQLApache Cassandra or Kafka Cloud and hybrid cloud deployments According to a 2021 performance assessment by RedHat, the growing use of hybrid cloud environments by enterprises is placing increased pressure on cloud-native storage, for which Presto, "the fastest distributed query engine available today" is ideal. 1 Moving workloads from an on-premise environment to a cloud or hybrid cloud infrastructure has many benefits, including increased performance and scalability. Presto's architecture makes it a strong choice for such deployments because it can be launched in a few minutes without additional provisioning, configuration or tuning. Machine learning (ML) Presto helps engineers prepare data and perform feature engineering and extraction in a highly efficient way that ensures it is ready for machine learning (ML). Its number of connectors, SQL engine and querying capabilities make it ideal for engineers seeking fast, easy access to large volumes of data. Additionally, Presto has tools designed specifically for ML functions such as aggregation, which allow data scientists to train support vector machine (SVM) classifiers and regressors to address supervised learning problems. Reporting Presto allows for data to be queried from multiple sources generating a single, easily accessible report or dashboard for BI purposes. Presto is simple and easy enough to use that analysts can conduct queries and create reports without the help of engineers. Analytics Presto enables analysts to conduct queries on both structured and unstructured data directly on a data lake without going through a data transformation process. Data preparation The process of gathering and preparing data can be costly and inefficient. Data scientists can spend hours each day collecting and preparing data before it can even be analyzed. Presto automates this process with speed and accuracy so data scientists and engineers can focus more of their time on higher value tasks. Related solutions Database software and solutions Use IBM database solutions to meet various workload needs across the hybrid cloud. Explore database solutions Cloud-native database with IBM Db2 Explore IBM Db2, a relational database that provides high performance, scalability and reliability for storing and managing structured data. It is available as SaaS on IBM Cloud or for self-hosting. Discover Db2 Data and analytics consulting services Unlock the value of enterprise data with IBM Consulting, building an insight-driven organization that delivers business advantage. Discover analytics services Structure your data to power faster low-latency transactions, continuous availability and hardened security. IBM® Db2® is an AI-powered database that accelerates decision making, reins in costs by using a single engine and promotes security for your business data. Discover IBM Db2 Explore database solutions Presto, o base de datos Presto (PrestoDB), es un motor de consultas SQL distribuido y de código abierto que puede consultar grandes conjuntos de datos de diferentes fuentes de datos, lo que permite a las empresas abordar los problemas de datos a gran escala. Presto ofrece a organizaciones de todos los tamaños una forma rápida y eficaz de analizar big data procedente de diversas fuentes, incluidos los sistemas en las instalaciones y la nube. También ayuda a las empresas a consultar petabytes de datos utilizando sus capacidades SQL actuales, sin tener que aprender un nuevo lenguaje. En la actualidad, Presto se utiliza principalmente para ejecutar consultas en Hadoop y otros proveedores de almacenamiento de datos comunes, lo que permite a los usuarios gestionar varios lenguajes de consulta e interfaces a bases de datos y almacenamiento. En la era digital, el análisis de big data se está convirtiendo rápidamente en una competencia fundamental para las empresas, independientemente de su tamaño o sector. La capacidad de recopilar, almacenar y analizar grandes cantidades de datos relacionados con los procesos empresariales, las preferencias de los clientes y las tendencias del mercado es extremadamente valiosa. La principal importancia de Presto para la analítica de datos es su capacidad para analizar los datos independientemente de dónde estén almacenados y sin tener que moverlos primero a un sistema más estructurado, como un almacén de datos o un data lake. Descubra ideas y noticias de expertos sobre IA, nube y mucho más en el boletín semanal Think. Presto se ha convertido en una herramienta popular para científicos e ingenieros de datos que trabajan con múltiples lenguajes de consulta, bases de datos aisladas y diferentes tipos de almacenamiento. Sus capacidades de alto rendimiento permiten a los usuarios consultar grandes volúmenes de datos en tiempo real, independientemente de dónde se encuentren los datos, utilizando una sencilla interfaz ANSI SQL. La velocidad y el rendimiento de Presto para ejecutar consultas en grandes volúmenes de datos lo han convertido en una herramienta indispensable para algunas de las empresas más grandes del mundo, como Facebook, Airbnb, Netflix, Microsoft, Apple (iOS) y AWS (Athena y Amazon 3). La arquitectura Presto es única, ya que está diseñada para consultar datos sin importar dónde se almacenen los datos, lo que lo hace más escalable y eficiente que otras soluciones similares. Las consultas de Presto permiten a los ingenieros utilizar los datos sin tener que moverlos físicamente de una ubicación a otra. Se trata de una capacidad importante, ya que las organizaciones se enfrentan a una cantidad cada vez mayor de datos que necesitan almacenar y analizar. Presto se creó para que los científicos e ingenieros de datos pudieran consultar de forma interactiva grandes cantidades de datos, independientemente de la fuente o el tipo de almacenamiento. Como Presto no almacena datos, sino que se comunica con una base de datos independiente para sus consultas, es más flexible que sus competidores y puede ampliar o reducir rápidamente las consultas en función de las necesidades cambiantes de la organización. Según un documento técnico de IBM, Presto, optimizado para las cargas de trabajo de inteligencia empresarial (BI), puede ayudar a las empresas a optimizar los precios de sus almacenes de datos y a reducir los costes hasta un 50 %. Estas son algunas de las principales ventajas de utilizar un flujo de trabajo de Presto: Menores costes: a medida que aumenta el tamaño de los almacenes de datos y la cantidad de usuarios que realizan consultas, no es raro que las empresas vean cómo sus costos aumentan rápidamente. Sin embargo, Presto está optimizado para grandes cantidades de consultas pequeñas, lo que facilita la consulta de cualquier cantidad de datos a la vez que mantiene los costes bajos. Además, como Presto es de código abierto, su implantación no conlleva gastos, lo que puede suponer un ahorro considerable para las empresas que deseen procesar grandes volúmenes de datos. Mayor escalabilidad: es habitual que los ingenieros configuren varios motores y lenguajes en un único sistema de almacenamiento de data lake, lo que puede hacer necesario volver a implantar la plataforma en el futuro y limitar la escalabilidad de su solución. Con Presto, todas las consultas se realizan utilizando el lenguaje y la interfaz universal ANSI SQL, lo que hace que la redefinición de la plataforma sea redundante. Además, Presto se puede utilizar para cantidades pequeñas y grandes de datos y escalar fácilmente de uno o dos usuarios a miles. Presto implementa múltiples motores de computación con dialectos SQL únicos y API, lo que lo convierte en una herramienta ideal para escalar cargas de trabajo que podrían ser demasiado complejas y requerir mucho tiempo para que los equipos de ingenieros y científicos de datos se manejen. Mejor rendimiento: aunque muchos motores de consulta que ejecutan SQL en Hadoop están restringidos en su rendimiento de computación porque están diseñados para escribir sus resultados en el disco, el modelo en memoria distribuido de Presto permite ejecutar grandes cantidades de consultas interactivas a la vez contra grandes conjuntos de datos. Tras un diseño clásico de procesamiento masivo paralelo (MPP), Presto programa tantas consultas como sea posible en un solo nodo de trabajo y utiliza la transmisión en secuencias en memoria para aumentar aún más sus velocidades de procesamiento. Ejecutar tareas en memoria hace que escribir y leer desde el disco entre etapas sea redundante y acorta el tiempo de cada ejecución de consulta, haciendo de Presto una opción de latencia inferior a la de sus competidores. Flexibilidad mejorada: Presto utiliza un modelo listo para usar en todas sus fuentes de datos, incluidas Cassandra, Kafka, MySQL, el sistema de archivos distribuidos de Hadoop (HDFS), PostgreSQL y otras, lo que hace que consultarlas sea más rápido y fácil que con otras herramientas comparables que carecen de esta funcionalidad. Además, la arquitectura flexible de Presto significa que no está restringida a un único proveedor, pero se ejecuta en la mayoría de las distribuciones de Hadoop, lo que lo convierte en una de las herramientas más portátiles disponibles. Aunque Presto no es la única opción de SQL en Hadoop disponible para desarrolladores e ingenieros de datos, su arquitectura única que mantiene la funcionalidad de consulta separada del almacenamiento de datos lo convierte en una de las más flexibles. A diferencia de otras herramientas, Presto separa el motor de consulta del almacenamiento de datos y utiliza conectores para comunicarse entre ellos. Esta funcionalidad añadida ofrece a los ingenieros más flexibilidad que otras herramientas a la hora de construir soluciones con Presto. Presto utiliza un sistema de administración de bases de datos MPP con un nodo coordinador que funciona en conjunto con otros nodos. Un ecosistema Presto se compone de tres tipos de servidores, un servidor coordinador, un servidor de trabajo y un servidor de administrador de recursos. Coordinador: un coordinador se considera el "cerebro" de una instalación de Presto. Es responsable de algunas de las tareas más importantes, como analizar declaraciones, planificar consultas y gestionar nodos de trabajo de Presto. En última instancia, es responsable de recuperar los datos de los nodos de trabajo y entregar los resultados al cliente. Trabajador: el trabajador es responsable de recopilar datos de los nodos de trabajo y garantizar el intercambio fluido de datos entre él y los conectores. Gestor de recursos: el gestor de recursos recopila datos de todos los nodos coordinadores y trabajadores y crea una vista global, o un "clúster Presto". Cuando el servidor SQL coordinador Presto recibe una consulta SQL de un usuario, lo primero que hace es utilizar una consulta personalizada para analizar, planificar y programar un plan distribuido en los otros nodos. La API REST de Presto se utiliza para enviar sentencias de consulta para su ejecución en un servidor y recuperar los resultados para el cliente. Presto admite significados ANSI SQL estándar, incluidas uniones, consultas, subconsultas y agregaciones. Una vez compilada la consulta, Presto analiza la solicitud en diferentes etapas entre los nodos de trabajo. Dado que Presto se basó en el concepto de abstracción de datos, es extensible a cualquier fuente de datos y puede consultar fácilmente fuentes de datos, como lagos de datos, almacenes de datos y bases de datos relacionales. La abstracción de datos es un proceso de programación que permite almacenar y manipular los datos de manera más eficiente al separar su representación de su almacenamiento físico. Esta abstracción permite que el motor de consultas se centre exclusivamente en los aspectos de los datos que son relevantes para su consulta. Mediante el proceso de abstracción de datos, los datos se consultan dondequiera que se almacenen, en lugar de una vez que se hayan trasladado a otro sistema de análisis. Inicialmente desarrollado en Facebook para ejecutar consultas interactivas en un enorme almacén de datos Apache Hadoop, los desarrolladores de Presto siempre lo imaginaban como software de código abierto y buscaban que fuera gratuito para uso comercial, para que cualquiera pudiera utilizarlo para el análisis de datos y la gestión de datos. En 2013, fue de código abierto en GitHub para que cualquiera lo descargue con la licencia de Apache Software. En 2019, tres de los miembros originales del equipo de desarrollo de Presto abandonaron el proyecto y fundaron una "bifurcación" de Presto conocida como Presto Software Foundation, o más comúnmente, prestosql. La base de Linux y otras comunidades de código abierto ofrecen webinars y formación sobre Presto en inglés y otros idiomas para ingenieros y desarrolladores que buscan obtener certificación. Estos foros también son un buen lugar para conocer las novedades de Presto. Presto permite a las organizaciones consultar repositorios de datos a gran escala y bases de datos NoSQL de forma rápida y eficiente para diversos fines empresariales. Estos son algunos de sus casos prácticos más comunes: Consultas específicas Presto permite una exploración rápida de datos e informes sencillos para diversos fines empresariales. Mediante el uso de conectores populares Presto, como Hive, MongoDB o Cassandra, los usuarios pueden consultar los datos que les interesen y obtener resultados en segundos. Con su velocidad y flexibilidad, Presto permite a los usuarios iterar y seguir explorando conjuntos de datos, independientemente de dónde residan. Estos son algunos de los repositorios de datos más utilizados a los que Presto puede conectarse: BigQueryHDFSCloud StorageSQL en la nube para MySQLApache Cassandra o Kafka Implementaciones de nube y nube híbrida Según una evaluación de rendimiento de 2021 realizada por RedHat, el uso cada vez mayor de entornos de nube híbrida por parte de las empresas está ejerciendo una mayor presión sobre el almacenamiento nativo de la nube, para lo cual Presto, "el motor de consultas distribuido más rápido disponible en la actualidad", es ideal 1. Mover cargas de trabajo de un entorno local a una infraestructura de nube o de nube híbrida tiene muchos beneficios, incluido un mayor rendimiento y escalabilidad. La arquitectura de Presto lo convierte en una buena opción para este tipo de implementaciones porque se puede iniciar en unos minutos sin necesidad de aprovisionamiento, configuración o ajustes adicionales. Machine learning (ML) Presto ayuda a los ingenieros a preparar los datos y a aplicar la ingeniería y extracción de características de una manera altamente eficiente que garantiza que estén listos para machine learning (ML). Su número de conectores, motor SQL y capacidades de consulta lo hacen ideal para los ingenieros que buscan un acceso rápido y fácil a grandes volúmenes de datos. Además, Presto cuenta con herramientas diseñadas específicamente para funciones de aprendizaje automático, como la agregación, que permiten a los científicos de datos entrenar clasificadores y regresores de máquinas vectoriales de soporte (SVM) para abordar problemas de aprendizaje supervisado. Informes Presto permite que los datos se consulten desde múltiples fuentes generando un único informe o panel de control de fácil acceso para fines de BI. Presto es lo suficientemente sencillo y fácil de usar como para que los analistas puedan realizar consultas y crear informes sin ayuda de ingenieros. Herramientas de análisis Presto permite a los analistas realizar consultas sobre datos estructurados y no estructurados directamente en un lago de datos sin pasar por un proceso de transformación de datos. Preparación de datos El proceso de recopilación y preparación de datos puede ser costoso e ineficiente. Los científicos de datos pueden dedicar horas cada día a recopilar y preparar datos antes de que se puedan analizar. Presto automatiza este proceso con velocidad y precisión para que los científicos e ingenieros de datos puedan dedicar más tiempo a tareas más importantes. Presto, atau basis data Presto (PrestoDB), adalah mesin kueri SQL terdistribusi sumber terbuka yang dapat mengkueri kumpulan data besar dari berbagai sumber data, yang memungkinkan perusahaan untuk mengatasi masalah data dalam skala besar. Presto memberi organisasi dari semua ukuran cara yang cepat dan efisien untuk menganalisis big data dari berbagai sumber termasuk sistem on premises dan cloud. Ini juga membantu bisnis melakukan kueri data berukuran petabyte dengan menggunakan kemampuan SQL yang ada saat ini, tanpa harus mempelajari bahasa baru. Presto saat ini umumnya digunakan untuk menjalankan kueri pada Hadoop dan penyedia penyimpanan data umum lainnya, memungkinkan pengguna untuk mengelola banyak bahasa kueri dan antarmuka ke database dan penyimpanan. Di era digital, analitik big data dengan cepat menjadi kompetensi inti bagi perusahaan terlepas dari ukuran atau industrinya. Kemampuan untuk mengumpulkan, menyimpan, dan menganalisis sejumlah besar data yang berkaitan dengan proses bisnis, preferensi pelanggan, dan tren pasar sangat berharga. Pentingnya Presto untuk analitik data adalah kemampuannya untuk menganalisis data di mana pun data disimpan, dan tanpa harus pindah ke sistem yang lebih terstruktur terlebih dahulu, seperti gudang data atau danau data. Temukan insight dan berita yang dikurasi oleh para pakar tentang AI, cloud, dan lainnya di Buletin Think mingguan. Presto telah menjadi alat yang populer bagi para ilmuwan data dan insinyur yang berurusan dengan berbagai bahasa kueri, database dengan silo, dan berbagai jenis penyimpanan. Kemampuan kinerjanya yang tinggi memungkinkan pengguna untuk meminta data dalam volume besar secara real-time, di mana pun data berada, menggunakan antarmuka ANSI SQL yang sederhana. Kecepatan dan kinerja Presto dalam mengeksekusi kueri pada volume data yang besar telah menjadikannya alat yang sangat diperlukan untuk beberapa perusahaan terbesar di dunia termasuk Facebook, Airbnb, Netflix, Microsoft, Apple (iOS) dan AWS (Athena dan Amazon s3). Arsitektur Presto unik karena dibuat untuk melakukan kueri data di mana pun data tersebut disimpan, sehingga membuatnya lebih dapat diskalakan dan efisien daripada solusi serupa lainnya. Kueri presto memungkinkan para engineer untuk menggunakan data tanpa harus memindahkannya secara fisik dari satu lokasi ke lokasi lain. Ini adalah kemampuan yang penting untuk dimiliki karena organisasi berurusan dengan jumlah data yang terus meningkat yang perlu mereka simpan dan analisis. Presto dibangun untuk memberdayakan para ilmuwan data dan teknisi untuk secara interaktif melakukan kueri sejumlah besar data terlepas dari sumber atau jenis penyimpanannya. Karena Presto tidak menyimpan data, melainkan berkomunikasi dengan database terpisah untuk kuerinya, Presto lebih fleksibel daripada pesaingnya dan dapat menskalakan kueri naik atau turun dengan cepat berdasarkan kebutuhan organisasi yang bergeser. Menurut whitepaper IBM, Presto, yang dioptimalkan untuk beban kerja intelijen bisnis (BI), dapat membantu perusahaan mengoptimalkan harga gudang data mereka dan mengurangi biaya hingga 50 persen. Berikut ini beberapa manfaat utama menggunakan alur kerja Presto: Biaya yang lebih rendah: Seiring bertambahnya ukuran gudang data dan jumlah pengguna yang melakukan kueri, tidak jarang perusahaan melihat biaya mereka meningkat dengan cepat. Presto, bagaimanapun, dioptimalkan untuk sejumlah besar kueri kecil, membuatnya mudah untuk melakukan kueri sejumlah data sambil juga menekan biaya. Ditambah lagi, karena Presto adalah sumber terbuka, tidak ada biaya yang terkait dengan penerapannya, yang dapat menghasilkan penghematan yang signifikan bagi perusahaan yang ingin memproses volume data yang besar. Peningkatan skalabilitas: Sudah umum bagi para insinyur untuk menyiapkan beberapa mesin dan bahasa pada satu sistem penyimpanan data lake, yang dapat membuatnya perlu untuk melakukan platform ulang di masa mendatang dan membatasi skalabilitas solusi mereka. Dengan Presto, semua kueri dilakukan menggunakan bahasa dan antarmuka ANSI SQL universal, membuat platform ulang menjadi berlebihan. Selain itu, Presto dapat digunakan untuk data dalam jumlah kecil dan besar dan mudah ditingkatkan dari satu atau dua pengguna menjadi ribuan. PrestoDB menerapkan beberapa mesin komputasi dengan dialek SQL yang unik dan API, menjadikannya alat yang ideal untuk menskalakan beban kerja yang mungkin terlalu rumit dan memakan waktu untuk ditangani oleh tim insinyur dan ilmuwan data. Performa yang lebih baik: Meskipun banyak mesin kueri yang menjalankan SQL di Hadoop dibatasi dalam performa komputasinya karena dibuat untuk menulis hasilnya ke disk, model in-memory terdistribusi dari Presto memungkinkannya untuk menjalankan kueri interaktif dalam jumlah besar sekaligus terhadap kumpulan data yang besar. Mengikuti desain pemrosesan paralel secara masif (MPP) klasik, Presto menjadwalkan sebanyak mungkin kueri pada satu node pekerja dan menggunakan pengacakan streaming dalam memori untuk meningkatkan kecepatan pemrosesan. Menjalankan tugas dalam memori membuat penulisan dan pembacaan dari disk di antara tahap-tahap menjadi berlebihan dan mempersingkat waktu eksekusi setiap kueri, sehingga Presto menjadi opsi latensi yang lebih rendah daripada pesaingnya. Fleksibilitas yang lebih baik: Presto menggunakan model plug-and-play untuk semua sumber datanya termasuk Cassandra, Kafka, MySQL, sistem file terdistribusi Hadoop (HDFS), PostgreSQL, dan lainnya, membuat kueri pada semua sumber data ini lebih cepat dan mudah dibandingkan dengan alat lain yang sebanding yang tidak memiliki fungsionalitas ini. Selain itu, arsitektur Presto yang fleksibel berarti tidak terbatas pada satu vendor saja, tetapi dapat berjalan di sebagian besar distribusi Hadoop, menjadikannya salah satu alat yang paling portabel yang tersedia. Meskipun Presto bukan satu-satunya opsi SQL-on-Hadoop yang tersedia untuk developer dan data engineer, arsitektur uniknya yang memisahkan fungsionalitas kueri dari penyimpanan data menjadikannya salah satu yang paling fleksibel. Tidak seperti alat lain, Presto memisahkan mesin kueri dari penyimpanan data dan menggunakan konektor untuk berkomunikasi di antara keduanya. Fungsionalitas tambahan ini memberikan para insinyur fleksibilitas yang lebih besar daripada alat lain dalam cara mereka membangun solusi menggunakan Presto. Presto menggunakan sistem manajemen database MPP dengan satu node koordinator yang bekerja bersama-sama dengan node lainnya. Ekosistem Presto terdiri dari tiga jenis server, yaitu server koordinator, server pekerja, dan server pengelola sumber daya. Koordinator: Seorang koordinator dianggap sebagai "otak" dari instalasi Presto. Koordinator bertanggung jawab untuk beberapa tugas yang paling penting termasuk penguraian pernyataan, merencanakan kueri dan mengelola node pekerja Presto. Pada akhirnya, ia bertanggung jawab untuk mengambil data dari node pekerja dan mengirimkan hasil ke klien. Pekerja: Pekerja bertanggung jawab untuk mengumpulkan data dari node pekerja dan memastikan kelancaran pertukaran data antara dirinya dan konektor. Resource manager: Resource manager mengumpulkan data dari semua node koordinator dan pekerja dan membuat tampilan global, atau "Presto cluster". Ketika SQL server koordinator Presto menerima SQL query dari pengguna, hal pertama yang dilakukannya adalah menggunakan kueri khusus untuk mengurai, merencanakan, dan menjadwalkan rencana terdistribusi di seluruh node lainnya. Presto REST API digunakan untuk mengirimkan pernyataan kueri untuk dieksekusi di server dan mengambil hasilnya bagi klien. Presto mendukung makna ANSI SQL standar, termasuk penggabungan, kueri, sub-kueri, dan agregasi. Setelah menyusun kueri, Presto menguraikan permintaan ke dalam beberapa tahap yang berbeda di antara node pekerja. Karena Presto dibangun di atas konsep abstraksi data, Presto dapat diperluas ke sumber data apa pun dan dapat dengan mudah melakukan kueri ke sumber data seperti data lake, gudang data, dan database relasional. Abstraksi data adalah proses pemrograman yang memungkinkan data disimpan dan dimanipulasi secara lebih efisien dengan memisahkan representasinya dari penyimpanan fisik. Abstraksi ini memungkinkan mesin kueri untuk fokus secara eksklusif pada aspek data yang relevan dengan kuerinya. Dengan menggunakan proses abstraksi data, data dilakukan kueri di mana pun data tersebut disimpan, bukan setelah data tersebut dipindahkan ke sistem analitik lain. Awalnya dikembangkan di Facebook untuk menjalankan kueri interaktif pada gudang data Apache Hadoop yang sangat besar, para pengembang Presto selalu membayangkannya sebagai perangkat lunak sumber terbuka dan berusaha membuatnya gratis untuk penggunaan komersial sehingga siapa pun dapat menggunakannya untuk analisis data dan manajemen data. Pada tahun 2013, aplikasi ini tersedia secara sumber terbuka di GitHub untuk diunduh oleh siapa saja di bawah lisensi Perangkat Lunak Apache. Pada tahun 2019, tiga anggota asli tim pengembangan Presto meninggalkan proyek dan mendirikan "fork" Presto yang dikenal sebagai Presto Software Foundation, atau lebih umum, sebagai prestosql. Yayasan Linux dan komunitas sumber terbuka lainnya menawarkan webinar dan pelatihan tentang Presto dalam bahasa Inggris dan bahasa lainnya untuk para insinyur dan pengembang yang ingin mendapatkan sertifikasi. Forum-forum ini juga merupakan tempat yang baik untuk dikunjungi untuk mengetahui apa yang baru di Presto. Presto memungkinkan organisasi untuk melakukan kueri terhadap repositori data berskala besar dan database NoSQL dengan cepat dan efisien untuk berbagai tujuan bisnis. Berikut adalah beberapa contoh penggunaan yang paling umum: Kueri ad-hoc Presto memungkinkan eksplorasi data yang cepat dan pelaporan langsung untuk berbagai tujuan bisnis. Menggunakan konektor Presto populer, seperti Hive, MongoDB atau Cassandra, pengguna dapat menanyakan data yang mereka minati dan mendapatkan hasil dalam hitungan detik. Dengan kecepatan dan fleksibilitasnya, Presto memberdayakan pengguna untuk mengulangi dan mengeksplorasi lebih lanjut kumpulan data, di mana pun mereka berada. Berikut ini adalah beberapa repositori data yang paling banyak digunakan yang dapat dihubungkan dengan Presto: BigQueryHDFSPenyimpanan CloudCloud SQL untuk MySQLApache Cassandra atau Kafka Penerapan cloud dan cloud hybrid Menurut penilaian kinerja tahun 2021 oleh RedHat, meningkatnya penggunaan lingkungan cloud hybrid oleh perusahaan memberikan tekanan yang meningkat pada penyimpanan cloud native, di mana Presto, "mesin kueri terdistribusi tercepat yang tersedia saat ini" sangat ideal. 1 Memindahkan beban kerja dari lingkungan lokal ke infrastruktur cloud atau cloud hybrid memiliki banyak manfaat, termasuk peningkatan performa dan skalabilitas. Arsitektur Presto membuatnya pilihan yang kuat untuk penerapan seperti itu karena dapat diluncurkan dalam beberapa menit tanpa penyediaan, konfigurasi atau penyetelan tambahan. Machine Learning (ML) Presto membantu para insinyur menyiapkan data dan melakukan rekayasa dan ekstraksi fitur dengan cara yang sangat efisien yang memastikan data siap untuk machine learning (ML). Jumlah konektor, mesin SQL dan kemampuan kueri menjadikannya ideal bagi para insinyur yang mencari akses cepat dan mudah ke volume data yang besar. Selain itu, Presto memiliki alat yang dirancang khusus untuk fungsi ML seperti agregasi, yang memungkinkan para ilmuwan data untuk melatih pengklasifikasi dan regressor mesin vektor pendukung (SVM) untuk mengatasi masalah pembelajaran yang diawasi. Pelaporan Presto memungkinkan data untuk ditanyakan dari berbagai sumber yang menghasilkan satu laporan atau dasbor yang mudah diakses untuk tujuan BI. Presto sederhana dan cukup mudah digunakan sehingga analis dapat melakukan kueri dan membuat laporan tanpa bantuan teknisi. Analytics Presto memungkinkan analis untuk melakukan kueri pada data terstruktur dan tidak terstruktur secara langsung pada data lake tanpa melalui proses transformasi data. Persiapan data Proses pengumpulan dan penyiapan data bisa jadi mahal dan tidak efisien. Ilmuwan data dapat menghabiskan waktu berjam-jam setiap harinya untuk mengumpulkan dan menyiapkan data sebelum data tersebut dianalisis. Presto mengotomatiskan proses ini dengan kecepatan dan akurasi sehingga para ilmuwan data dan insinyur dapat memfokuskan lebih banyak waktu mereka pada tugas-tugas yang bernilai lebih tinggi. Solusi terkait Perangkat lunak dan solusi database Gunakan solusi database IBM untuk memenuhi berbagai kebutuhan beban kerja di hybrid cloud. Jelajahi solusi database Database cloud native dengan IBM Db2 Jelajahi IBM Db2, database relasional yang menghadirkan kinerja tinggi, skalabilitas, dan keandalan untuk menyimpan dan mengelola data terstruktur. Database ini tersedia sebagai SaaS di IBM Cloud atau untuk hosting mandiri. Jelajahi Db2 Layanan konsultasi data dan analitik Buka nilai data perusahaan dengan IBM Consulting, membangun organisasi berbasis insight yang memberikan keuntungan bisnis. Temukan layanan analitik Presto , 가 . ANSI SQL . Presto Facebook, Airbnb, Netflix, Microsoft, Apple(iOS) AWS(Athena Amazon s3) 가 . Presto 가 . Presto . 가 . Presto 가 . Presto , . IBM , (BI) Presto 가 50% . Presto 가 . : 가 가 가 가 . Presto . Presto . : 가 , . Presto 가 ANSI SQL . Presto . Presto SQL API . : Hadoop SQL , Presto . (MPP) Presto 가 . 가 Presto . : Presto Cassandra, Kafka, MySQL, Hadoop (HDFS), PostgreSQL . Presto Hadoop 가 . Presto 가 SQL-on-Hadoop , 가 . Presto . 가 가 Presto . Presto、またはPrestoデータベース（PrestoDB）は、さまざまなデータソースから大規模なデータセットを照会できるオープンソースの分散SQL Queryエンジンなので、企業はデータの大規模な問題にも対処できます。 組織の規模を問わず、オンプレミス・システムやクラウドなどのさまざまなソースからのビッグデータを迅速かつ効率的に分析できるのがPrestoの強みです。また、新しい言語を学習することなく、既存のSQL機能を使用して、ペタバイト規模のデータを照会するのにも役立ちます。 Prestoは、Hadoopやその他の一般的なデータ・ストレージ・プロバイダーにおいて、クエリーの実行に最もよく使用されています。ユーザーは、複数のクエリー言語とデータベースやストレージへのインターフェイスを管理できます。 デジタル時代において、ビッグデータ分析は、規模や業界にかかわらず、急速に企業にとってのコア・コンピテンシーになりつつあります。ビジネス・プロセス、顧客の嗜好、市場動向に関する大量のデータを収集、保存、分析できる能力には、大きな価値があります。データ分析におけるPrestoの強みは、データが保存されている場所にかかわらず、データウェアハウスやデータレイクなどの、より構造化されたシステムに最初にデータを移動する必要がなく、データを分析できる点にあります。「The DX Leaders」は日本語でお届けするニュースレターです。AI活用のグローバル・トレンドや日本の市場動向を踏まえたDX、生成AIの最新情報を毎月お届けします。 Prestoは、複数のクエリ言語、サイロ化されたデータベース、さまざまな種類のストレージを扱うデータサイエンティストやエンジニアに人気のツールです。ユーザーは、データの場所にかかわらず、シンプルなANSI SQLインターフェイスを使用して、大量のデータをリアルタイムに照会できます。Prestoは、大量のデータに対してクエリを実行できるスピードとパフォーマンスにより、Facebook、Airbnb、Netflix、Microsoft、Apple（iOS）、AWS（AthenaとAmazon s3）など、世界をリードする企業に不可欠なツールとなっています。 Prestoのアーキテクチャーは、データの保存場所にかかわらず、データを照会できるように設計されています。したがって、他の類似のソリューションよりもスケーラブルで効率的です。Prestoのクエリを使用することで、エンジニアはデータを物理的に移動することなく、データを活用できるようになります。これは、組織が保存、分析する必要がある、増え続けるデータを処理するうえできわめて重要な機能です。 Prestoは、データサイエンティストやエンジニアが、ソースやストレージの種類にかかわらず、膨大なデータをインタラクティブに照会できるように設計されています。Prestoでは、データの保管はしないものの、別のデータベースと通信して照会するので、競合製品よりも柔軟性が高く、組織のニーズの変化に基づいたクエリのスケールアップもスケールダウンも迅速になります。IBMのホワイトペーパーによると、Prestoをビジネス・インテリジェンス（BI）ワークロードに最適化すれば、企業にはデータ・ウェアハウスの価格設定を最適化してコストを最大50%削減する助けとなります。 Prestoのワークフローを導入する主なメリットをご紹介します。 コストの削減：データ・ウェアハウスのサイズとクエリを実行するユーザー数が増加するにつれて、企業のコストは急速に増加します。Prestoは、大量の小規模なクエリ向けに最適化されているため、コストを抑えながらも、あらゆる量のデータを簡単に照会することができます。また、Prestoはオープンソースであるため、デプロイに関連する料金はかかりません。膨大なデータを処理する必要がある企業にとっては、大きな節約につながる可能性があります。 拡張性の向上：エンジニアが1つのデータレイク・ストレージ・システムに複数のエンジンと言語をセットアップするのはよくあることですが、将来的にはプラットフォームの再構築が必要になってソリューションの拡張性が制限されることがあります。Prestoでは、すべてのクエリがユニバーサルANSI SQL言語とインターフェイスを使用して実行されるため、プラットフォームの再構築をしなくてもよくなります。さらに、Prestoはデータの量が小さくても大きくても使用できるので、1～2ユーザーから数千ユーザーへのスケールアップも簡単です。独自のSQLダイアレクトとAPIにより多数のコンピューティング・エンジンをデプロイするPrestoは、エンジニアやデータサイエンティストのチームが処理するには複雑すぎて時間のかかりそうなワークロードをスケーリングするための理想的なツールとなります。 パフォーマンスの向上：Hadoop上でSQLを実行する多くのクエリエンジンは、結果をディスクに書き込むように設計されているため、計算パフォーマンスに制約があります。Prestoの分散インメモリ・モデルでは、大規模なデータセットに対して、一度に大量のインタラクティブ・クエリを実行することができます。Prestoは、古典的な超並列処理（MPP）設計に従って、単一のWorkerノード上でできる限り多くのクエリをスケジュールし、インメモリ・ストリーミング・シャッフルを使用して、処理速度をさらに高めます。メモリ内でタスクを実行すると、ステージ間のディスクへの書き込みと読み取りが冗長になり、各クエリの実行時間が短縮されるため、競合ツールよりも待機時間が短くなります。 柔軟性の向上：Prestoでは、そのすべてのデータソースで、Cassandra、Kafka、MySQL、Hadoop分散ファイルシステム（HDFS）、PostgreSQLなどのプラグアンドプレイ・モデルを使用しているので、こうした機能がない他の競合ツールに比べて、データソース間のクエリがより高速で容易になります。また、Prestoはその柔軟なアーキテクチャーにより、単一のベンダーに限定されることなく、ほとんどのHadoopディストリビューションで稼働するので、入手可能な中で最もポータブルなツールの1つとなっています。 Prestoは、開発者やデータエンジニアが利用できる唯一のSQL-on-Hadoopではありませんが、クエリ機能をデータ・ストレージから切り離す独自のアーキテクチャーにより、最も柔軟なオプションの1つです。他のツールとは異なり、クエリ・エンジンをデータ・ストレージから分離し、コネクターを使用してクエリ・エンジン間の通信を行います。この機能により、エンジニアはPrestoを使用して、より柔軟にソリューションを構築することができます。 Prestoは、他のノードと連携して動作する、1つのCoordinatorノードを備えたMPPデータベース管理システムを使用します。Prestoのエコシステムは、Coordinatorサーバー、Workerサーバー、Resource Managerサーバーの3つのサーバー・タイプで構成されています。 Coordinator：コーディネーターは、Prestoをインストールする際の「頭脳」です。ステートメントの解析、クエリの計画、Presto Workerノードの管理など、最も重要なタスクのいくつかを担当します。最終的には、Workerノードからデータを取得し、結果をクライアントに配信します。 Worker：Workerは、Workerノードからデータを収集し、Worker自身とコネクタ間でデータがスムーズに交換されるようにします。 Resource Manager：Resource Managerでは、すべてのCoordinatorノードとWorkerノードからデータを収集してグローバル・ビュー（「Prestoクラスター」）を作成します。 Presto Coordinator SQL Serverサーバーがユーザーから SQL Queryを受け取ると、最初にカスタム・クエリーを使用して解析・計画し、他のノード全体に分散プランをスケジュールします。Presto REST APIは、サーバー上で実行するクエリー・ステートメントを送信し、クライアントの結果を取得するために使用されます。Prestoは、結合、クエリー、サブクエリー、集計などの標準的なANSI SQLのミーニングをサポートしています。クエリーをコンパイルすると、PrestoはリクエストをWorkerノード間のさまざまなステージに解析します。 Prestoはデータ抽象化の概念に基づいて構築されているため、あらゆるデータソースに拡張可能で、データレイク、データ・ウェアハウス、リレーショナル・データベースなどのデータソースを簡単に照会できます。データ抽象化は、データの表現を物理的なストレージから分離することで、より効率的にデータを保存・操作できるようにするプログラミング・プロセスです。この抽象化により、クエリ・エンジンはクエリに関連するデータの側面のみに焦点を当てることができます。データの抽象化プロセスを使用すると、データが別の分析システムに移動されてからではなく、保存されている場所であればどこでもデータを照会できます。 当初は、巨大なApache Hadoopデータウェアハウスでインタラクティブ・クエリを実行するためにFacebookで開発されましたが、Prestoの開発者は常にオープンソース・ソフトウェアとして構想し、誰でもデータ分析やデータ管理に使用できるよう、無料で商用利用できるようにすることを目指していました。2013年には、GitHubでオープンソース化され、Apacheソフトウェア・ライセンスの下で、誰でもダウンロードできるようになりました。2019年、Presto開発チームのオリジナル・メンバーのうち3人がプロジェクトを離れ、Presto Software Foundation（より一般的には、prestosqlとして知られる）Prestoの「フォーク」を設立しました。 Linux Foundationやその他のオープンソース・コミュニティーは、認定資格の取得を目指すエンジニアや開発者向けに、Prestoに関するウェビナーやトレーニングを英語やその他の言語で提供しています。これらのフォーラムは、Prestoの最新情報を調べるのに最適な場所です。 Prestoを使用すると、組織はさまざまなビジネス目的で大規模なデータ・リポジトリやNoSQLデータベースを迅速かつ効率的に照会できます。ここでは、その最も一般的なユースケースをいくつかご紹介します。 アドホック・クエリ さまざまなビジネス目的での、迅速なデータ探索と簡単なレポート作成が可能になります。Hive、MongoDB、Cassandraなどの一般的なPrestoコネクターを使用することで、ユーザーは興味のあるデータを照会し、数秒で結果を得ることができます。Prestoのスピードと柔軟性により、ユーザーはどこにいても、データセットを反復処理して、さらに探索できるようになります。 Prestoが接続できる、最も広く使用されているデータ・リポジトリの一部は次のとおりです。 BigQueryHDFSクラウドストレージCloud SQL for MySQLApache CassandraまたはKafka クラウドとハイブリッドクラウドの導入 RedHatが実施した2021年のパフォーマンスアセスメントによると、企業によるハイブリッドクラウド環境の利用拡大により、クラウドネイティブ・ストレージへの圧力が高まっており、「現在利用可能な最速の分散クエリ・エンジン」であるPrestoが理想的とされています。1ワークロードをオンプレミス環境からクラウドかハイブリッドクラウド・インフラストラクチャーへ移行することには、パフォーマンスや拡張性の向上など、多くのメリットがあります。Prestoのアーキテクチャーは、そうしたデプロイメントでの強力な選択肢となります。プロビジョニング、構成、チューニングをまた行うことなしに数分で起動できるためです。 機械学習（ML） Prestoは、エンジニアがデータを準備し、特徴量エンジニアリングと抽出を極めて効率的な方法で実行し、機械学習（ML）の準備を確実に整える助けとなります。コネクターの数、SQLエンジン、クエリ機能を備えているため、大量のデータにもすぐに簡単にアクセスしようとしているエンジニアに最適です。さらに、Prestoには集計などのML関数に特化して設計されているツールがあるので、データサイエンティストは、サポート・ベクター・マシン（SVM）の分類と回帰をトレーニングすることで、教師あり学習の問題に対処できます。 レポート Prestoを使用すると、複数のソースからデータを照会して、BI目的で簡単にアクセスできる単一のレポートまたはダッシュボードを作成できます。Prestoはシンプルで使いやすいため、アナリストはエンジニアの助けを借りずにクエリを実行し、レポートを作成できます。 分析 アナリストは、データ変換をすることなく、構造化データと非構造化データの両方に対して、データレイク上で直接クエリを実行できます。 データ準備 データの収集と準備にはコストがかかり、非効率的である可能性があります。データサイエンティストは、データを分析する前のデータの収集と準備に、毎日何時間も費やすことがあります。Prestoにより、このプロセスを迅速かつ正確に自動化できるため、データサイエンティストやエンジニアはより価値の高いタスクにより多くの時間を割くことができます。 Presto oder Presto-Datenbank (PrestoDB) ist eine verteilte Open-Source-SQL-Query-Engine, die große Datensätze aus verschiedenen Datenquellen abfragen kann, sodass Unternehmen Datenprobleme im großen Maßstab lösen können. Presto bietet Unternehmen jeder Größe eine schnelle und effiziente Möglichkeit zur Analyse von Big Data aus verschiedenen Quellen, einschließlich lokaler Systeme und der Cloud. Es hilft Unternehmen außerdem dabei, Petabytes an Daten mithilfe ihrer aktuellen SQL-Funktionen abzufragen, ohne eine neue Sprache erlernen zu müssen. Heute wird Presto am häufigsten für die Ausführung von Abfragen auf Hadoop und anderen gängigen Datenspeicheranbietern verwendet und ermöglicht den Benutzern die Verwaltung mehrerer Abfragesprachen und Schnittstellen zu Datenbanken und Speichern. Im digitalen Zeitalter entwickelt sich Big Data Analytics schnell zu einer Kernkompetenz für Unternehmen, unabhängig von Größe oder Branche. Die Fähigkeit, große Datenmengen über Geschäftsprozesse, Kundenpräferenzen und Markttrends zu sammeln, zu speichern und zu analysieren, ist äußerst wertvoll. Presto kann Daten unabhängig von ihrem Speicherort und ohne vorheriges Verschieben in ein strukturierteres System (beispielsweise ein Data Warehouse oder einen Data Lake) analysieren – das macht Presto so bedeutsam. Entdecken Sie von Experten kuratierte Erkenntnisse und Neuigkeiten zu KI, Cloud und mehr im wöchentlichen Newsletter Think. Abonnieren Sie noch heute Presto hat sich als beliebtes Tool für Data Scientists und Entwicklungsteams etabliert, die mit mehreren Abfragesprachen, isolierten Datenbanken und verschiedenen Speichertypen arbeiten. Durch die leistungsstarken Funktionen können mithilfe einer einfachen ANSI-SQL-Schnittstelle große Datenmengen in Echtzeit abgefragt werden, ganz egal, wo die Daten gespeichert sind. Prestos Geschwindigkeit und Leistung beim Abfragen großer Datenmengen haben es zu einem unverzichtbaren Tool für einige der größten Unternehmen der Welt gemacht, darunter Facebook, Airbnb, Netflix, Microsoft, Apple (iOS) und AWS (Athena und Amazon s3). Die Presto-Architektur ist einzigartig, da Daten unabhängig von ihrem Speicherort abgefragt werden können. Dadurch ist sie skalierbarer und effizienter als andere ähnliche Lösungen. Mit Presto-Abfragen können Daten genutzt werden, ohne sie zuvor an einen anderen Speicherort verschieben zu müssen. Dies ist ein wichtiger Aspekt, da die Menge an zu speichernden und zu analysierenden Daten ständig zunimmt. Presto wurde entwickelt, damit Data Scientists und Entwicklungsteams große Datenmengen unabhängig von der Quelle oder Art des Speichers interaktiv abfragen können. Presto selbst speichert keine Daten, sondern kommuniziert bei seinen Abfragen mit einer separaten Datenbank. Dadurch ist es flexibler als die Konkurrenz und kann Abfragen gemäß den sich ändernden Anforderungen des Unternehmens schnell erhöhen oder zurückschrauben. Laut einem Whitepaper von IBM kann Presto, optimiert für Business Intelligence (BI)-Workloads, die Preise für Data Warehouses optimieren und die Kosten um bis zu 50 Prozent senken. Hier sind einige der wichtigsten Vorteile für die Verwendung eines Presto-Workflows: Kostengünstiger: Da Data Warehouses immer größer werden und immer mehr Benutzer Abfragen durchführen, können die Kosten in kurzer Zeit rasant steigen. Presto ist jedoch für große Mengen kleiner Abfragen optimiert. Es lässt sich also jede beliebige Datenmenge unkompliziert abfragen, ohne die Kosten in die Höhe zu treiben. Da Presto eine Open-Source-Lösung ist, ist die Bereitstellung gebührenfrei, was bei der Verarbeitung großer Datenmengen erhebliche Einsparungen bedeuten kann. Erhöhte Skalierbarkeit:Es kommt häufig vor, dass Ingenieure mehrere Engines und Sprachen auf einem einzigen Data-Lake-Speichersystem einrichten, was zu einem späteren Zeitpunkt eine neue Plattform erforderlich machen und die Skalierbarkeit ihrer Lösung einschränken kann. Bei Presto werden alle Abfragen über die universelle ANSI-SQL-Sprache und -Schnittstelle durchgeführt, so dass keine neue Plattform nötig ist. Außerdem funktioniert Presto mit kleinen Datenmengen ebenso gut wie mit großen und lässt sich leicht von einem oder zwei Benutzern auf Tausende aufstocken. Presto stellt mehrere Verarbeitungs-Engines mit eigenen SQL-Dialekten und APIs bereit. Das macht es zum idealen Tool zur Skalierung von Workloads, die für Teams aus Ingenieuren und Data Scientists zu komplex und zeitaufwändig sein könnten. Leistungsstärker: Viele Abfrage-Engines, die SQL auf Hadoop ausführen, können ihre Verarbeitungsleistung nur eingeschränkt nutzen, weil sie ihre Ergebnisse auf die Festplatte schreiben. Im Gegensatz dazu ermöglicht das dezentrale In-Memory-Modell von Presto die gleichzeitige Ausführung großer Mengen interaktiver Abfragen an großen Datensätzen. Nach einem klassischen MPP-Aufbau (Massive Parallel Processing) terminiert Presto möglichst viele Abfragen auf einem einzigen Arbeitsknoten und verwendet In-Memory-Streaming-Shuffle, um die Verarbeitungsgeschwindigkeit noch weiter zu erhöhen. Durch die speicherinterne Ausführung von Tasks wird das Schreiben und Lesen von der Festplatte zwischen den einzelnen Phasen überflüssig und die einzelnen Abfragen werden schneller. Daher hat Presto eine geringere Latenzzeit als die Konkurrenz. Flexibler: Presto arbeitet bei allen seinen Datenquellen, einschließlich Cassandra, Kafka, MySQL, Hadoop Distributed File System (HDFS), PostgreSQL und andere, mit einem Plug-and-Play-Modell. Das beschleunigt und vereinfacht die Abfragen im Vergleich zu anderen Tools ohne diese Funktionalität. Außerdem ist Presto durch seine flexible Architektur nicht auf einen einzelnen Anbieter beschränkt, sondern kann auf den meisten Hadoop-Distributionen ausgeführt werden – weshalb es eines der portabelsten Tools ist. Presto ist zwar nicht die einzige SQL-on-Hadoop-Option für Entwicklungsteams und Data Scientists, aber durch seine einzigartige Architektur, die die Abfragefunktionalität von der Datenspeicherung trennt, eines der flexibelsten. Im Gegensatz zu anderen Tools isoliert Presto die Abfrage-Engine von der Datenspeicherung und kommuniziert mit ihnen über Konnektoren. Diese zusätzliche Funktionalität bietet bei der Erstellung von Lösungen mit Presto mehr Flexibilität als andere Tools. Presto verwendet ein MPP-Datenbank-Verwaltungssystem mit einem Koordinatorknoten, der mit anderen Knoten zusammenarbeitet. Ein Presto-Ökosystem besteht aus drei Arten von Servern: einem Koordinator-Server, einem Worker-Server und einem Ressourcenmanager-Server. Koordinator: Gilt als „Gehirn" einer Presto-Installation. Er ist für einige besonders wichtige Aufgaben zuständig, darunter das Parsen von Anweisungen, die Planung von Abfragen und die Verwaltung von Presto-Arbeitsknoten. Letztendlich ist es seine Aufgabe, Daten von den Worker-Knoten abzurufen und die Ergebnisse an den Client zu liefern. Worker: Der Worker ist dafür verantwortlich, Daten von den Workerknoten zu sammeln und den reibungslosen Datenaustausch zwischen ihm und den Konnektoren sicherzustellen. Resource Manager: Der

Resource Manager sammelt Daten von allen Koordinator- und Worker-Knoten und erstellt eine globale Ansicht oder ein „Presto-Cluster“. Wenn der SQL-Server des Presto-Koordinators eine SQL-Abfrage erhält, wird zuerst mit einer benutzerdefinierten Abfrage ein dezentraler Plan für die anderen Knoten geparst, geplant und terminiert. Über die Presto-REST-API werden Abfragen zur Ausführung auf einem Server eingereicht und Ergebnisse für den Client abgerufen. Presto unterstützt Standard-ANSI-SQL-Aussagen, einschließlich Joins, Abfragen, Unterabfragen und Aggregationen. Nach dem Kompilieren der Abfrage parst Presto die Anfrage in verschiedene Phasen zwischen den Worker-Knoten. Da Presto auf dem Konzept der Datenabstraktion basiert, kann es auf jede Datenquelle erweitert werden und ohne großen Aufwand auch Data Lakes, Data Warehouses und relationale Datenbanken abfragen. Datenabstraktion bedeutet, dass die Darstellung der Daten von ihrem physischen Speicher getrennt wird. Dies erfolgt beim Programmieren, um Daten effizienter zu speichern und zu bearbeiten. Durch die Abstraktion kann sich die Abfrage-Engine ausschließlich auf die für ihre Abfrage relevanten Aspekte der Daten konzentrieren. Mit der Datenabstraktion werden die Daten dort abgefragt, wo sie gespeichert sind, müssen also nicht erst in ein anderes Analysesystem verschoben werden. Ursprünglich wurde Presto bei Facebook entwickelt, um interaktive Abfragen in einem riesigen Hadoop-Data-Warehouse unter Apache auszuführen. Für die Entwickler von Presto war es immer eine Open-Source-Software. Es sollte kostenfrei kommerziell nutzbar sein, damit jeder damit Daten analysieren und verwalten konnte.Im Jahr 2013 wurde es als Open-Source-Version auf GitHub bereitgestellt und konnte unter der Apache-Softwarelizenz heruntergeladen werden. Im Jahr 2019 verließen drei der ursprünglichen Mitglieder des Presto-Entwicklungsteams das Projekt und gründeten einen „Fork“ von Presto, bekannt als Presto Software Foundation oder auch „prestosql“. Die Linux Foundation und andere Open-Source-Communitys bieten Webinare und Schulungen zu Presto in Englisch und anderen Sprachen für Personen an, die eine Zertifizierung anstreben. Diese Foren sind auch ein guter Ort für alle Neuigkeiten rund um Presto. Mit Presto können Unternehmen große Datenspeicher und NoSQL-Datenbanken schnell und effizient für eine Vielzahl von Geschäftszwecken abfragen. Hier sind einige der häufigsten Anwendungsfälle: Ad-hoc-Abfragen Presto ermöglicht die schnelle Untersuchung von Daten und eine unkomplizierte Berichterstellung für verschiedenste Geschäftszwecke. Mit beliebten Presto-Konnektoren wie Hive, MongoDB oder Cassandra können Benutzer Daten von Interesse abfragen und erhalten in Sekundenschnelle ein Ergebnis. Dank seiner Schnelligkeit und Flexibilität können Datensätze mit Presto abgeglichen und weiter untersucht werden, unabhängig davon, wo sie gespeichert sind. Hier sind einige der am häufigsten verwendeten Daten-Repositorys, zu denen Presto eine Verbindung herstellen kann: BigQueryHDFSCloudspeicherCloud-SQL für MySQLApache Cassandra oder Kafka Cloud- und Hybrid-Cloud-Bereitstellungen Laut einer Leistungsbewertung durch RedHat von 2021 steigt der Druck auf cloudnative Speicher, weil es immer mehr Hybrid-Cloud-Umgebungen gibt. Presto ist hier als „schnellste heute verfügbare dezentrale Abfrage-Engine“ ideal geeignet. 1 Die Verlagerung von Workloads aus einem lokalen System in eine Cloud- oder Hybrid-Cloud-Infrastruktur bietet viele Vorteile, darunter höhere Leistung und Skalierbarkeit. Die Architektur von Presto macht es zu einer guten Wahl für solche Bereitstellungen, da es in wenigen Minuten ohne zusätzliche Installation, Konfiguration oder Feineinstellung ausführbar ist. Maschinelles Lernen (ML) Presto ermöglicht eine hoch effiziente Datenaufbereitung sowie Erstellung und Extraktion von Merkmalen, um Daten für maschinelles Lernen (ML) vorzubereiten. Aufgrund seiner zahlreichen Konnektoren, der SQL-Engine und der Abfragefunktionen ist es ideal, wenn schnell und einfach auf große Datenmengen zugegriffen werden muss. Darüber hinaus verfügt Presto über speziell für ML-Funktionen wie Aggregation entwickelte Tools. Damit können SVM-Klassifikatoren und -Regressoren (Support Vector Machine) trainiert werden, um Probleme des überwachten Lernens zu beheben. Berichterstellung Mit Presto können Daten aus mehreren Quellen abgefragt werden. Es generiert daraus einzelne gut zugängliche Berichte oder Dashboards für BI-Zwecke. Presto ist so einfach und benutzerfreundlich, dass Abfragen oder die Berichterstellung auch ohne Hilfe von Fachkräften möglich sind. Analysen Presto ermöglicht es Analysten, Abfragen sowohl für strukturierte als auch für unstrukturierte Daten direkt in einem Data Lake durchzuführen, ohne einen Datenkonvertierungsprozess durchlaufen zu müssen. Datenaufbereitung Daten zu erfassen und aufzubereiten kann kostspielig und ineffizient sein. Data Scientists könnten jeden Tag Stunden damit zubringen, Daten zu sammeln und aufzubereiten, bevor sie überhaupt analysierbar sind. Presto automatisiert diesen Prozess schnell und präzise, damit Data Scientists und Entwicklungsteams mehr Zeit für wertschöpfendere Aufgaben haben. Weiterführende Lösungen Datenbanksoftware und -lösungen Mit IBM Datenbanklösungen können Sie verschiedene Workload-Anforderungen in der Hybrid Cloud erfüllen. Datenbank-Lösungen erkunden Cloudnativ-Datenbank mit IBM Db2 Erfahren Sie mehr über IBM Db2: eine relationale Datenbank, die hohe Leistung, Skalierbarkeit und Zuverlässigkeit für das Speichern und Verwaltung strukturierter Daten bietet. Die Lösung ist als SaaS in der IBM Cloud oder als Self-Hosting-Option verfügbar. Db2 entdecken Beratungsservices für Daten und Analysen Erschließen Sie den Wert von Unternehmensdaten mit IBM Consulting und bauen Sie ein erkenntnisorientiertes Unternehmen auf, das Ihnen geschäftliche Vorteile verschafft. Analyse-Services entdecken O Presto, ou banco de dados Presto (PrestoDB), é um mecanismo de consultas de SQL distribuído, de código aberto, capaz de consultar grandes volumes de dados de diferentes fontes, permitindo que as empresas lidem com problemas de dados em larga escala. Ele oferece às organizações de todos os portes uma maneira rápida e eficiente de analisar big data de várias fontes, incluindo sistemas no local e na nuvem. Também ajuda as empresas a consultarem petabytes de dados usando seus recursos atuais de SQL, sem precisar aprender uma nova linguagem. Atualmente, o Presto é mais utilizado para executar consultas em Hadoop e em outros provedores comuns de armazenamento de dados, permitindo que os usuários gerenciem vários idiomas de consulta e interfaces com bancos de dados e com o armazenamento. Na era digital, análise de big data vem se tornando rapidamente uma competência essencial para empresas, independentemente do porte ou do setor. A capacidade de coletar, armazenar e analisar grandes volumes de dados relacionados a processos de negócios, preferências dos clientes e tendências de mercado tem um valor imenso. A principal vantagem do Presto para a análise de dados é conseguir analisar os dados independentemente de onde estejam armazenados, sem precisar transferi-los previamente para um sistema mais estruturado, como um data warehouse ou um data lake. Descubra insights selecionados por especialistas e notícias sobre IA, nuvem e outros assuntos no boletim informativo semanal Think. O Presto se tornou uma ferramenta popular para cientistas e engenheiros de dados que lidam com várias linguagens de consulta, bancos de dados isolados e diferentes tipos de armazenamento. Seus recursos de alto desempenho permitem que os usuários consultem grandes volumes de dados em tempo real, independentemente de onde os dados estejam localizados, usando uma interface ANSI SQL simples. A velocidade e o desempenho do Presto na realização de consultas em grandes volumes de dados tornaram-no uma ferramenta indispensável para algumas das maiores empresas do mundo, como Facebook, Airbnb, Netflix, Microsoft, Apple (iOS) e AWS (Athena e Amazon s3). A arquitetura Presto é única, pois foi criada para consultar dados, não importa onde eles estejam sendo armazenados, tornando-a mais escalável e eficiente do que outras soluções semelhantes. As consultas Presto permitem que os engenheiros usem dados sem precisar migrá-los fisicamente de um local para outro. Este é um recurso importante, à medida que as organizações lidam com uma quantidade cada vez maior de dados que elas precisam armazenar e analisar. O Presto foi criado para permitir que cientistas de dados e engenheiros consultem grandes volumes de dados de forma interativa, independentemente da origem ou do tipo de armazenamento. Como o Presto não armazena dados, mas se comunica com um banco de dados separado para executar suas consultas, ele oferece mais flexibilidade que os concorrentes e consegue aumentar ou reduzir a escala das consultas rapidamente conforme as necessidades variáveis da organização. Segundo um whitepaper da IBM, o Presto, otimizado para cargas de trabalho de business intelligence (BI), pode ajudar empresas a otimizar os custos dos seus data warehouses e a reduzir despesas em até 50%. Estes são alguns dos principais benefícios de usar um fluxo de trabalho Presto: Custos mais baixos: à medida que o tamanho dos data warehouses e o número de usuários que realizam consultas aumentam, não é incomum que as empresas vejam seus custos aumentarem rapidamente. O Presto, no entanto, é otimizado para grandes quantidades de pequenas consultas, o que facilita a consulta de qualquer quantidade de dados e, ao mesmo tempo, mantém os custos baixos. Além disso, como o Presto é de código aberto, não há taxas associadas à sua implementação, o que pode resultar em uma economia significativa para as empresas que desejam processar grandes volumes de dados. Escalabilidade aumentada: É comum que engenheiros configurem múltiplos mecanismos e linguagens sobre um único sistema de armazenamento de data lake, o que pode exigir reconfigurações futuras e limitar a escalabilidade da solução. Com o Presto, todas as consultas são feitas com a linguagem e interface SQL ANSI universal, tornando redundante novas plataformas. Além disso, o Presto pode ser usado tanto com pequenos quanto com grandes volumes de dados, sendo facilmente escalável de um ou dois usuários para milhares. O Presto implementa diversos mecanismos de processamento com dialetos SQL exclusivos e APIs, tornando-se uma ferramenta ideal para escalar cargas de trabalho que poderiam ser complexas e demoradas demais para equipes de engenheiros e cientistas de dados. Melhor desempenho: embora muitos mecanismos de consulta que executam SQL no Hadoop tenham desempenho computacional restrito porque foram criados para gravar seus resultados em disco, o modelo distribuído em memória do Presto permite que ele realize grandes quantidades de consultas interativas de uma só vez em grandes conjuntos de dados. Seguindo um modelo clássico de processamento paralelo massivo (MPP), o Presto agenda o máximo de consultas possível em um único nó de trabalho e usa o streaming aleatório na memória para aumentar ainda mais suas velocidades de processamento. A execução de tarefas na memória torna redundantes a gravação e a leitura do disco entre os estágios e reduz o tempo de execução de cada consulta, tornando o Presto uma opção de menor latência do que seus concorrentes. Maior flexibilidade: o Presto usa um modelo plug-and-play para todas as suas fontes de dados, incluindo Cassandra, Kafka, MySQL, sistema de arquivos distribuído do Hadoop (HDFS), PostgreSQL e outras, tornando as consultas entre elas mais rápidas e mais simples do que com outras ferramentas semelhantes que não oferecem essa funcionalidade. Além disso, a arquitetura flexível do Presto não se restringe a um único fornecedor e roda na maioria das distribuições Hadoop, o que o torna uma das ferramentas mais portáteis disponíveis. Embora o Presto não seja a única opção de SQL-on-Hadoop disponível para desenvolvedores e engenheiros de dados, sua arquitetura exclusiva que mantém a funcionalidade de consulta separada do armazenamento de dados o torna um dos mais flexíveis. Ao contrário de outras ferramentas, o Presto separa o mecanismo de consulta do armazenamento de dados e usa conectores para a comunicação entre eles. Essa funcionalidade adicional oferece aos engenheiros mais flexibilidade do que outras ferramentas na forma como eles constroem soluções usando o Presto. O Presto usa um sistema de gerenciamento de banco de dados MPP com um nó coordenador que funciona em conjunto com outros nós. Um ecossistema Presto é composto por três tipos de servidores, um servidor coordenador, um servidor de trabalho e um servidor de resource manager. Coordenador: o servidor coordenador é considerado o "cérebro" da instalação do Presto. Ele é responsável por algumas das tarefas mais críticas, incluindo a análise de declarações, o planejamento de consultas e gerenciamento dos nós de trabalho do Presto. Por fim, ele é responsável por recuperar os dados dos nós de trabalho e entregar os resultados ao cliente. De trabalho: o servidor de trabalho é responsável por coletar os dados dos nós de trabalho e garantir a troca tranquila de dados entre ele e os conectores. Gerenciador de recursos: o gerenciador de recursos coleta dados de todos os nós coordenadores e nós de trabalho e cria uma visão global, ou um “cluster Presto”. Quando o servidor SQL coordenador do Presto recebe uma consulta SQL de um usuário, ele começa usando uma consulta personalizada para interpretar, planejar e agendar a execução distribuída entre os demais nós. A API REST do Presto é usada para enviar instruções de consulta para execução no servidor e recuperar os resultados para o cliente. O Presto é compatível com os significados padrão do SQL ANSI, incluindo junções, consultas, subconsultas e agregações. Após compilar a consulta, o Presto divide a requisição em diferentes estágios entre os nós de trabalho. Como o Presto foi criado com base no conceito de abstração de dados, ele é extensível a qualquer fonte de dados e pode facilmente consultar fontes de dados, como data lakes, data warehouses e bancos de dados relacionais. A abstração de dados é um processo de programação que permite que os dados sejam armazenados e manipulados de maneira mais eficiente, separando sua representação do seu armazenamento físico. Essa abstração permite que o mecanismo de consulta se concentre exclusivamente nos aspectos dos dados relevantes para sua consulta. Por meio do processo de abstração de dados, os dados são consultados onde quer que estejam sendo armazenados, e não depois de terem sido migrados para outro sistema de análise de dados. Inicialmente desenvolvido no Facebook para realizar consultas interativas em um enorme data warehouse Apache Hadoop, os desenvolvedores do Presto sempre o imaginaram como um software de código aberto e procuraram torná-lo gratuito para uso comercial para que qualquer pessoa pudesse usá-lo para análise e gerenciamento de dados. Em 2013, ele foi disponibilizado de código aberto no GitHub para que qualquer pessoa pudesse baixar sob a licença do Apache Software. Em 2019, três dos membros originais da equipe de desenvolvimento do Presto abandonaram o projeto e fundaram uma versão alternativa do Presto conhecida como Presto Software Foundation, ou mais comumente prestosql. A fundação Linux e outras comunidades de código aberto oferecem webinars e treinamentos sobre o Presto em inglês e outros idiomas para engenheiros e desenvolvedores que desejam obter certificação. Esses fóruns também são um bom lugar para descobrir as novidades do Presto. O Presto permite que as organizações consultem repositórios de dados de grande escala e bancos de dados NoSQL de forma rápida e eficiente para uma variedade de fins comerciais. Estes são alguns dos casos de uso mais comuns: Consultas ad-hoc O Presto permite a exploração rápida de dados e relatórios diretos para uma variedade de fins comerciais. Por meio de conectores Presto populares, como Hive, MongoDB ou Cassandra, os usuários podem consultar dados de interesse e obter resultados em segundos. Com sua velocidade e flexibilidade, o Presto possibilita os usuários a iterar e explorar ainda mais os conjuntos de dados, independentemente de onde estejam. Estes são alguns dos repositórios de dados mais usados aos quais o Presto pode se conectar: BigQueryHDFSArmazenamento em nuvemCloud SQL para MySQLApache Cassandra ou Kafka Implementações em nuvem e em nuvem híbrida Segundo uma avaliação de desempenho de 2021 da Red Hat, o uso crescente de ambientes de nuvem híbrida pelas empresas aumenta a pressão sobre o armazenamento nativo da nuvem; para isso, o Presto, “o mecanismo de consulta distribuído mais rápido disponível hoje”, é ideal. 1 Transferir cargas de trabalho de um ambiente local para uma infraestrutura de nuvem ou nuvem híbrida traz muitos benefícios, incluindo melhor desempenho e escalabilidade. A arquitetura do Presto o torna uma excelente opção para essas implementações, pois pode ser iniciado em poucos minutos, sem provisionamento, configuração ou ajuste adicionais. Aprendizado de máquina (ML) O Presto ajuda engenheiros a preparar dados e executar engenharia e extração de atributos de forma altamente eficiente, garantindo que eles estejam prontos para aprendizado de máquina (ML). Seu número de conectores, o mecanismo SQL e seus recursos de consulta tornam o Presto ideal para engenheiros que precisam de acesso rápido e fácil a grandes volumes de dados. Além disso, o Presto oferece ferramentas projetadas especificamente para funções de ML, como agregação, que permitem que cientistas de dados treinem classificadores e regressores de máquinas de vetores de suporte (SVM) para lidar com problemas de aprendizado supervisionado. Relatórios O Presto permite que os dados sejam consultados a partir de várias fontes e gera um único relatório ou dashboard facilmente acessível para fins de BI. O Presto é tão simples e fácil de usar que os analistas podem realizar consultas e criar relatórios sem a ajuda de engenheiros. Análise de dados O Presto permite que os analistas realizem consultas em dados estruturados e não estruturados diretamente em um data lake, sem passar por um processo de transformação de dados. Preparação de dados O processo de coletar e preparar dados pode ser caro e ineficiente. Os cientistas de dados podem passar horas todos os dias coletando e preparando dados antes mesmo de analisá-los. O Presto automatiza esse processo com rapidez e precisão, permitindo que cientistas de dados e engenheiros dediquem mais tempo a tarefas de maior valor. Soluções relacionadas Software e soluções de bancos de dados Use soluções de bancos de dados da IBM para atender a diversas necessidades de cargas de trabalho em toda a nuvem híbrida. Explore as soluções de banco de dados Banco de dados nativo da nuvem com o IBM Db2 Explore o IBM Db2, um banco de dados relacional que oferece alto desempenho, escalabilidade e confiabilidade para armazenar e gerenciar dados estruturados. Está disponível como SaaS no IBM Cloud ou para auto-hospedagem. Descubra o Db2 Serviços de consultoria de dados e análise de dados Libere o valor dos dados empresariais com a IBM® Consulting, construindo uma organização orientada por insights, que proporciona vantagem comercial. Conheça os serviços de análise de dados Presto, o base de datos Presto (PrestoDB), es un motor de código abierto SQL Query distribuido que puede consultar grandes conjuntos de datos de diferentes fuentes, lo que permite a las compañías abordar problemas de datos a escala. Presto ofrece a organizaciones de todos los tamaños una manera rápida y eficiente de analizar big data de diversas fuentes, incluidos los sistemas on-premise y la nube. También ayuda a las compañías a consultar petabytes de datos por medio de sus capacidades SQL actuales, sin tener que aprender un nuevo idioma. En la actualidad, Presto se usa más para ejecutar consultas en Hadoop y otros proveedores de almacenamiento de datos comunes, lo que permite a los usuarios administrar varios lenguajes de consulta e interfaces para bases de datos y almacenamiento. En la era digital, el analytics de big data se está convirtiendo rápidamente en una competencia fundamental para las empresas, independientemente de su tamaño o industria. La capacidad de recopilar, almacenar y analizar grandes cantidades de datos relacionados con los procesos empresariales, las preferencias de los clientes y las tendencias del mercado es extremadamente valiosa. La principal importancia de Presto para el analytics de datos es su capacidad de analizar los datos independientemente de dónde se almacenen los datos y sin tener que moverlos primero a un sistema más estructurado, como un almacén de datos o un lago de datos. Descubra insights y noticias de expertos sobre IA, la nube y mucho más en el boletín semanal Think. Presto se ha convertido en una herramienta popular para científicos e ingenieros de datos que trabajan con múltiples lenguajes de consulta, bases de datos aisladas y diferentes tipos de almacenamiento. Sus capacidades de alto rendimiento permiten a los usuarios consultar grandes volúmenes de datos en tiempo real, independientemente de dónde se encuentren los datos, utilizando una sencilla interfaz ANSI SQL. La velocidad y el rendimiento de Presto para ejecutar consultas en grandes volúmenes de datos lo han convertido en una herramienta indispensable para algunas de las empresas más grandes del mundo, como Facebook, Airbnb, Netflix, Microsoft, Apple (iOS) y AWS (Athena y Amazon 3). La arquitectura Presto es única, ya que está diseñada para consultar datos sin importar dónde se almacenen los datos, lo que lo hace más escalable y eficiente que otras soluciones similares. Las consultas de Presto permiten a los ingenieros utilizar los datos sin tener que moverlos físicamente de una ubicación a otra. Se trata de una capacidad importante, ya que las organizaciones se enfrentan a una cantidad cada vez mayor de datos que necesitan almacenar y analizar. Presto se creó para que los científicos e ingenieros de datos pudieran consultar de forma interactiva grandes cantidades de datos, independientemente de la fuente o el tipo de almacenamiento. Como Presto no almacena datos, sino que se comunica con una base de datos independiente para sus consultas, es más flexible que sus competidores y puede ampliar o reducir rápidamente las consultas en función de las necesidades cambiantes de la organización. Según un reporte técnico de IBM, Presto, optimizado para cargas de trabajo de business intelligence (BI), puede ayudar a las compañías a optimizar los precios de sus almacenes de datos y reducir los costos hasta en un 50 por ciento. Estos son algunos de los beneficios clave de utilizar un flujo de trabajo de Presto: Costos más bajosMayor escalabilidadMejor rendimientoFlexibilidad mejorada Costos más bajos a medida que aumenta el tamaño de los almacenes de datos y la cantidad de usuarios que realizan consultas, no es raro que las empresas vean cómo sus costos aumentan rápidamente. Sin embargo, Presto está optimizado para grandes cantidades de consultas pequeñas, lo que facilita la consulta de cualquier cantidad de datos a la vez que mantiene los costes bajos. Además, como Presto es de código abierto, su despliegue no conlleva gastos, lo que puede suponer un ahorro considerable para las empresas que deseen procesar grandes volúmenes de datos. Mayor escalabilidad Es común que los ingenieros configuren varios motores e idiomas en un único sistema de almacenamiento de lagos de datos, lo que puede hacer necesario volver a la plataforma en el futuro y limitar la escalabilidad de su solución. Con Presto, todas las consultas se realizan utilizando el lenguaje y la interfaz universal ANSI SQL, lo que hace que la redefinición de la plataforma sea redundante. Además, Presto se puede utilizar para cantidades pequeñas y grandes de datos y escalar fácilmente de uno o dos usuarios a miles. Presto despliega múltiples motores de computación con dialectos y API SQL únicos, lo que lo convierte en una herramienta ideal para escalar cargas de trabajo que podrían ser demasiado complejas y requerir mucho tiempo para que los equipos de ingenieros y científicos de datos se manejen. Mejor rendimiento Aunque muchos motores de consulta que ejecutan SQL en Hadoop están restringidos en su rendimiento de computación porque están diseñados para escribir sus resultados en el disco, el modelo en memoria distribuido de Presto permite ejecutar grandes cantidades de consultas interactivas a la vez contra grandes conjuntos de datos. Tras un diseño clásico de procesamiento masivo paralelo (MPP), Presto programa tantas consultas como sea posible en un solo nodo de trabajo y utiliza la transmisión en secuencias en memoria para aumentar aún más sus velocidades de procesamiento. Ejecutar tareas en memoria hace que escribir y leer desde el disco entre etapas sea redundante y acorta el tiempo de cada ejecución de consulta, haciendo de Presto una opción de latencia inferior a la de sus competidores. Flexibilidad mejorada Presto utiliza un modelo plug-and-play para todas sus fuentes de datos, como Cassandra, Kafka, MySQL, el sistema de archivos distribuido Hadoop (HDFS), PostgreSQL y otros, lo que hace que las consultas en ellos sean más rápidas y fáciles que con otras herramientas comparables que carecen de esta funcionalidad. Además, la arquitectura flexible de Presto significa que no está restringida a un único proveedor, pero se ejecuta en la mayoría de las distribuciones de Hadoop, lo que lo convierte en una de las herramientas más portátiles disponibles. Aunque Presto no es la única opción de SQL en Hadoop disponible para desarrolladores e ingenieros de datos, su arquitectura única que mantiene la funcionalidad de consulta separada del almacenamiento de datos lo convierte en una de las más flexibles. A diferencia de otras herramientas, Presto separa el motor de consulta del almacenamiento de datos y utiliza conectores para comunicarse entre ellos. Esta funcionalidad añadida ofrece a los ingenieros más flexibilidad que otras herramientas a la hora de construir soluciones con Presto. Presto utiliza un sistema de administración de bases de datos MPP con un nodo coordinador que funciona en conjunto con otros nodos. Un ecosistema Presto se compone de tres tipos de servidores, un servidor coordinador, un servidor de trabajo y un servidor de administrador de recursos. Coordinador Un coordinador se considera el “cerebro” de una instalación de Presto. Es responsable de algunas de las tareas más críticas, como analizar declaraciones, planificar consultas y gestionar nodos de trabajo de Presto. En última instancia, es responsable de recuperar los datos de los nodos de trabajo y entregar los resultados al cliente. Trabajador El trabajador es responsable de recopilar datos de los nodos de trabajo y garantizar el intercambio fluido de datos entre él y los conectores. Resource Manager El Resource Manager recopila datos de todos los nodos de coordinador y trabajador y crea una vista global, o un “cluster Presto”. Cuando el SQL Server coordinador Presto recibe una SQL Query de un usuario, lo primero que hace es utilizar una consulta personalizada para analizar, planificar y programar un plan distribuido en los otros nodos. La API REST de Presto se utiliza para enviar sentencias de consulta para su ejecución en un servidor y recuperar los resultados para el cliente. Presto admite significados ANSI SQL estándar, incluidas uniones, consultas, subconsultas y agregaciones. Una vez compilada la consulta, Presto analiza la solicitud en diferentes etapas entre los nodos de trabajo. Dado que Presto se basó en el concepto de abstracción de datos, es extensible a cualquier fuente de datos y puede consultar fácilmente fuentes de datos, como lagos de datos, almacenes de datos y bases de datos relacionales. La abstracción de datos es un proceso de programación que permite almacenar y manipular los datos de manera más eficiente al separar su representación de su almacenamiento físico. Esta abstracción permite que el motor de consultas se centre exclusivamente en los aspectos de los datos que son relevantes para su consulta. Mediante el proceso de abstracción de datos, los datos se consultan dondequiera que se almacenen, en lugar de una vez que se hayan trasladado a otro sistema de análisis. Inicialmente desarrollado en Facebook para ejecutar consultas interactivas en un enorme almacén de datos Apache Hadoop, los desarrolladores de Presto siempre lo imaginaban como software de código abierto y buscaban que fuera gratuito para uso comercial, para que cualquiera pudiera utilizarlo para el análisis de datos y la gestión de datos. En 2013, fue de código abierto en GitHub para que cualquiera lo descargue con la licencia de Apache Software. En 2019, tres de los miembros originales del equipo de desarrollo de Presto abandonaron el proyecto y fundaron una “bifurcación” de Presto conocida como Presto Software Foundation, o más comúnmente, prestosql. La fundación Linux y otras comunidades de código abierto ofrecen seminarios web y capacitación sobre Presto en inglés y otros idiomas para ingenieros y desarrolladores que buscan obtener la certificación. Estos foros también son un buen lugar para conocer las novedades de Presto. Presto permite a las organizaciones consultar repositorios de datos a gran escala y bases de datos NoSQL de forma rápida y eficiente para diversos fines empresariales. Estos son algunos de sus casos de uso más comunes: Consultas específicas Presto permite una exploración rápida de datos e informes sencillos para diversos fines empresariales. Mediante el uso de conectores populares Presto, como Hive, MongoDB o Cassandra, los usuarios pueden consultar los datos que les interesen y obtener resultados en segundos. Con su velocidad y flexibilidad, Presto permite a los usuarios iterar y seguir explorando conjuntos de datos, independientemente de dónde residan. Estos son algunos de los repositorios de datos más utilizados a los que Presto puede conectarse: BigQueryHDFSAlmacenamiento en la nubeCloud SQL para MySQLApache Cassandra o Kafka Despliegues de nube y nube híbrida Según una evaluación de rendimiento de 2021 realizada por RedHat, el uso cada vez mayor de entornos de nube híbrida por parte de las empresas está ejerciendo una mayor presión sobre el almacenamiento nativo de la nube, para lo cual Presto, “el motor de consultas distribuido más rápido disponible en la actualidad”, es ideal. 1 Mover cargas de trabajo de un entorno on-premises a una infraestructura de nube o de nube híbrida tiene muchos beneficios, incluido un mayor rendimiento y escalabilidad. La arquitectura de Presto lo convierte en una buena opción para este tipo de despliegues porque se puede iniciar en unos minutos sin aprovisionamiento, configuración o ajuste adicionales. Aprendizaje automático (ML) Presto ayuda a los ingenieros a preparar los datos y a realizar la ingeniería y extracción de características de una manera altamente eficiente que garantiza que estén listos para el machine learning (ML). Su número de conectores, motor SQL y capacidades de consulta lo hacen ideal para los ingenieros que buscan un acceso rápido y fácil a grandes volúmenes de datos. Además, Presto cuenta con herramientas diseñadas específicamente para funciones de aprendizaje automático, como la agregación, que permiten a los científicos de datos entrenar clasificadores y regresores de máquinas vectoriales de soporte (SVM) para abordar problemas de aprendizaje supervisado. Presentación de informes Presto permite que los datos se consulten desde múltiples fuentes generando un único informe o panel de fácil acceso para fines de BI. Presto es lo suficientemente sencillo y fácil de usar como para que los analistas puedan realizar consultas y crear informes sin ayuda de ingenieros. Analytics Presto permite a los analistas realizar consultas sobre datos estructurados y no estructurados directamente en un lago de datos sin pasar por un proceso de transformación de datos. Preparación de los datos El proceso de recopilación y preparación de datos puede ser costoso e ineficiente. Los científicos de datos pueden dedicar horas cada día a recopilar y preparar datos antes de que se puedan analizar. Presto automatiza este proceso con velocidad y precisión para que los científicos e ingenieros de datos puedan dedicar más tiempo a tareas de mayor valor.