Beyond Hadoop

snowflake

MODERN CLOUD DATA WAREHOUSING

CHAMPION GUIDES

What's inside

snowflake

How the hype and hope turned to headache

The benefits and liabilities that define Hadoop

When evaluating architectures for an enterprise big data platform and data analytics, Hadoop often makes the shortlist. Hadoop burst onto the data processing and analytics scene a decade ago with great fanfare. The promise of this open source, Java-based programming framework was threefold:

- Rewrite the economics of data analytics by supporting storage and processing of extremely large data sets on commodity hardware nodes.
- Help decision-makers conquer exploding volumes of big data.
- Deliver new business insights unattainable with traditional data warehousing and analytics solutions.

A GROWING BUT STILL COMPARATIVELY SMALL MARKET

Decades before Hadoop arrived, the traditional data warehouse emerged in the 1990s. As the data warehouse evolved, it created significant opportunities and high levels of demand the technology failed to meet. The pressure on the data warehouse mounted further in the early 2000s as the volume, velocity and variety of available data snowballed. Thus, the need for a better approach set the stage for the birth of Hadoop.

A huge amount of Hadoop hype emerged as organizations hoped it would remove the limitations of their overstretched, traditional data warehouses. Forrester Research sized the 2016 Hadoop market at $554 million, estimating that it will grow to $2.98 billion by 2021.[1] But these numbers are much smaller than the $203 billion that IDC projects for the big data and business analytics market for 2020.[2] Specific to enterprise data warehousing, Forrester Research sized its 2016 market at $16.96 billion, growing to an estimated $26.48 billion in 2021.[3]

Certainly, Hadoop has instilled a desire in companies to harness insight from data in new, imaginative ways. However, Hadoop presents several key challenges for organizations:

- High levels of customization required because Hadoop consists of multiple components that must be deployed and integrated by customers.
- Lack of specialized skills required to deploy, manage and use Hadoop.
- Constant need for support, only addressed at the cost of engaging independent, commercial support providers.
- Poor performance that limits Hadoop to batch processing and limited numbers of users.

How can enterprises get the processing potential of Hadoop and the best of traditional data warehousing, and still benefit from related emerging technologies?

This champion guide explores the options available to meet the needs of the modern enterprise, while outlining a path to simple, powerful and scalable solutions available today with modern data warehousing built for the cloud.

[1] "Big Data Solutions Forecast, 2016 to 2021," Jennifer Adams, Forrester Research, August 22, 2016. https://www.forrester.com/report/Forrester+Data+Big+Data+Management+Solutions+Forecast+2016+To+2021+Global/-/E-RES135913

[2] "Double-Digit Growth Forecast for the Worldwide Big Data and Business Analytics Market Through 2020 Led by Banking and Manufacturing Investments, According to IDC," press release, October 3, 2016. http://www.idc.com/getdoc.jsp?containerId=prUS41826116

[3] Source: "Big Data Solutions Forecast, 2016 to 2021," Jennifer Adams, Forrester Research, August 22, 2016. https://www.forrester.com/report/Forrester+Data+Big+Data+Management+Solutions+Forecast+2016+To+2021+Global/-/E-RES135913

Hadoop's inherent complexity

The challenges of NoSQL in a SQL world

Hadoop provides a framework for distributing data storage and processing across thousands of potential commodity servers. This framework is designed for scalability, flexibility and resiliency. The storage component, the Hadoop Distributed File System (HDFS), provides the important task of storing and replicating data. The processing components support multiple processing engines for running distributed tasks across multiple servers.

A RELIANCE ON SCARCE SKILLS

Hadoop is not a programming language. Rather, it's a framework. The original processing framework available with Hadoop, which is MapReduce, is a programming paradigm originated within Google for constructing custom distributed applications. Writing MapReduce programs requires specialized programming skills in distributed programming that remain relatively scarce.

Later improvements to Hadoop, in particular the YARN framework, have made it possible for processing engines other than MapReduce to be used within Hadoop. However, regardless of the processing engine used with Hadoop, the same fundamental challenge remains: Hadoop requires programming skills and expertise that are in scarce supply. In spite of many attempts to make Hadoop compatible with more widely available skills, programming in Hadoop still requires a rare combination of expertise in distributed programming and data management.

MEANWHILE, SQL PROGRAMMING SKILLS ARE WIDELY AVAILABLE

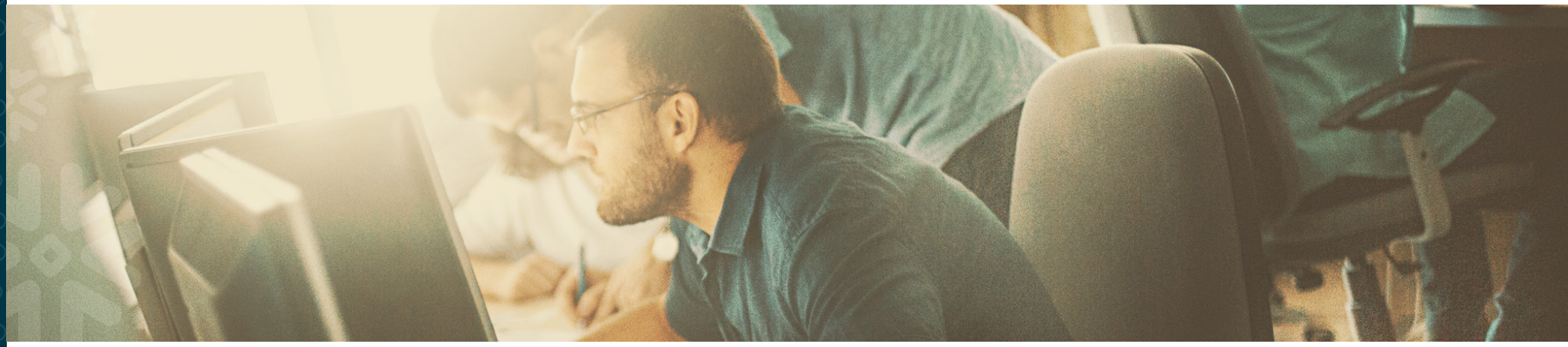
To query data, structured query language (SQL) is the established language of choice and the standard of the massive RDBMS market. Due to its decades-long prominence and broad support, SQL expertise is widely available. Survey results from research firm Evans Data Corporation and the developer community Stack Overflow put the estimated number of developers working with SQL today at seven million.[4] Similarly, a multitude of data integration and business intelligence tools integrate with SQL databases.

Over the years, there have been numerous efforts to support SQL within Hadoop. But these efforts have delivered fragmented and, sometimes, conflicting results. The first such approach, the Hive project, essentially translated SQL commands into MapReduce programs — an inefficient approach that fell short in terms of SQL compatibility and performance. Later efforts to create SQL engines to run within Hadoop produced a fragmented number of different approaches not designed for SQL relational processing.

[4] Source: Stack Overflow developer survey 2015; Evans Data Corporation Global Developer Population and Demographic Study, 2016 Volume 2. https://www.jetbrains.com/datagrip/how-developers-use-databases-today/

How Hadoop performs

Built for scale, not for speed.



ARCHITECTED FOR THROUGHPUT AT SCALE

Hadoop was designed to be a framework to store and process large volumes of data. It can indeed scale. Yahoo, the originator of Hadoop, is reported to have 600 petabytes stored in its Hadoop cluster. The ability to simply scale the capacity of a Hadoop cluster[5] by adding additional compute nodes has been an appealing capability for many organizations.

NOT DESIGNED FOR PERFORMANCE AND CONCURRENCY

However, fundamental design decisions embedded in the Hadoop architecture optimize for throughput and scale, but not for performance and efficiency. Yahoo illustrates this trade-off. The 600 petabytes of data in Hadoop at Yahoo require 40,000 compute nodes to store and process that data. The architecture of Hadoop requires that data be stored on the compute nodes in the cluster. This forces Hadoop clusters to be sized to accommodate the full data size even when the bulk of that data goes unused most of the time. The massive clusters required to store large amounts of data create significant overhead for processing, which slows performance. Additional architectural features further limit the performance of Hadoop.

Hadoop also has limited ability to support concurrent users and workloads. The advanced resource management required to support concurrent workloads with predictable performance are unavailable in Hadoop. As a result, Hadoop deployments typically support only a small number of concurrent workloads and users.

These shortcomings limit Hadoop to batch workloads, where performance is not critical. Organizations with workloads that require interactive, ad hoc processing and higher levels of concurrency have looked beyond Hadoop for more effective solutions.

[5] https://www.datanami.com/2015/10/12/inside-yahoos-super-sized-deep-learning-cluster/

Hadoop not ready for enterprise environments

Complexity creates significant overhead

The table below lists and describes the open source projects that provide the core functionality commonly referred to as Apache Hadoop version 2 and later. Other Hadoop components with additional functionality are also available.

HADOOP COMPONENT	FUNCTIONALITY
HADOOP DISTRIBUTION FILE SYSTEM (HDFS)	Framework for data storage
YARN	Cluster technology for resource management
MAPREDUCE 2	MapReduce libraries for computation and a job logger
APACHE PIG	Data flow programming language
APACHE HIVE	MapReduce based software for creating higher level, SQL-like queries (using HiveQL)
APACHE HCATALOG	Metadata abstraction layer for determining how and where data is physically stored
APACHE ZOOKEEPER	Tool for providing services to Hadoop servers (necessary for HBase installations)
APACHE OOZIE	Server-based workflow engine optimized for executing Hadoop jobs
APACHE KNOX	Single access point REST API gateway for interacting with Apache Hadoop clusters
APACHE SPARK	In-memory data processing engine for streaming and machine learning workloads used within Hadoop via YARN
APACHE PHOENIX	SQL relational database layer for Apache HBase
APACHE TEZ	Framework for building high performance batch and interactive data processing applications, coordinated by YARN
APACHE FALCON	Framework for simplifying and orchestrating data management and pipeline processing in Apache Hadoop
APACHE RANGER	Centralized security policy administration for authorization, auditing and data protection of Hadoop clusters

A MYRIAD OF MOVING PARTS

Hadoop is a broad collection of specialized open source projects that enable much more than storage and processing. Choosing, configuring, integrating and managing the components for a Hadoop project is significant. Individual open source projects under the Hadoop umbrella evolve and change rapidly. In addition, people with a strong understanding of the full ecosystem, and those with the skills to use and manage Hadoop, are in short supply.

HIGH-MAINTENANCE HADOOP ENVIRONMENTS

Each addition to a Hadoop environment adds to the complexity. Each component has its own development and release cycle from the Apache Software Foundation. This complex, high-maintenance burden is what drives Hadoop users to purchase support contracts from commercial Hadoop distributors. It also forces organizations to evaluate simpler options.

Delivering an enterprise-class solution based on Hadoop adds further overhead and complexity. Security, change control, data protection and monitoring are just a few of the enterprise requirements that demand specific expertise and effort to enable a Hadoop environment.

Break free from the complexity of Hadoop

Technology built for the cloud offers dramatic simplicity

While nearly one-third of enterprises have tried Hadoop, the relatively modest amount of revenue associated with the overall Hadoop[6] market suggests enterprises have not widely adopted this technology for big data processing. SQL-based enterprise data warehouses still dominate. As mentioned above, over the next few years, the data warehousing market will grow more than 12x in less time than will the Hadoop market.

The core promise of data warehousing remains central to what organizations need for their data: A standards-based, enterprise-class solution for high performance querying and reporting on data at scale. However, traditional on-premises data warehousing has not kept up with the exploding demands placed on data analytics. As a result, new architectures and solutions for data warehousing have emerged.

MODERN CLOUD DATA WAREHOUSES DELIVER THE BEST OF BOTH WORLDS

Modern cloud data warehousing presents a dramatically simpler but more power approach than both Hadoop and traditional on-premises or "cloud-washed" data warehouse solutions:[7]

- The ability to process and analyze large amounts of rapidly incoming data. Similar to how Hadoop operates, modern cloud data warehouses take this capability a step further by natively ingesting a broad range of data types, including traditional structured data such as CSV[8] files, as well as semi-structured data such as JSON,[9] with no transformation required.
- A robust SQL data processing environment. Modern cloud data warehouses enable this data processing with familiar and robust ANSI[10]–and ACID[11]–compliant SQL semantics known to millions of applications, tools, programmers and users.

Data warehousing built for the cloud can further free IT organizations from the complexity of both Hadoop and traditional warehousing with:

- Fully independent system resources: storage, compute and management, with dynamic scaling
- Zero-administration management with no "knobs" to turn or tune
- Automatic data distribution and metadata management
- Enterprise-class security, including automatic creation and handling of data encryption keys

» In short, data warehousing built for the cloud enables the big data processing that today's enterprises require, but with simplicity and ease unmatched by Hadoop or on-premises and cloud-washed traditional data warehouse solutions.

[6] "Forrester Sees Steady Growth for Big Data, Hadoop, and NoSQL," Alex Woodie, Datanami, September 16, 2016. https://www.datanami.com/2016/09/16/forrester-sees-steady-growth-big-data-Hadoop-nosql/

[7] Source: "Cloud-washed" refers to legacy data warehouse architectures that have been merely transferred to the cloud, instead of being re-architected for the cloud.

[8] Source: Comma-separated values

[9] JavaScript Object Notation

[10] American National Standards Institute

[11] ACID is a concept referring to a database system's four transaction properties: atomicity, consistency, isolation and durability.

The modern cloud data warehouse

Three components architected for the cloud

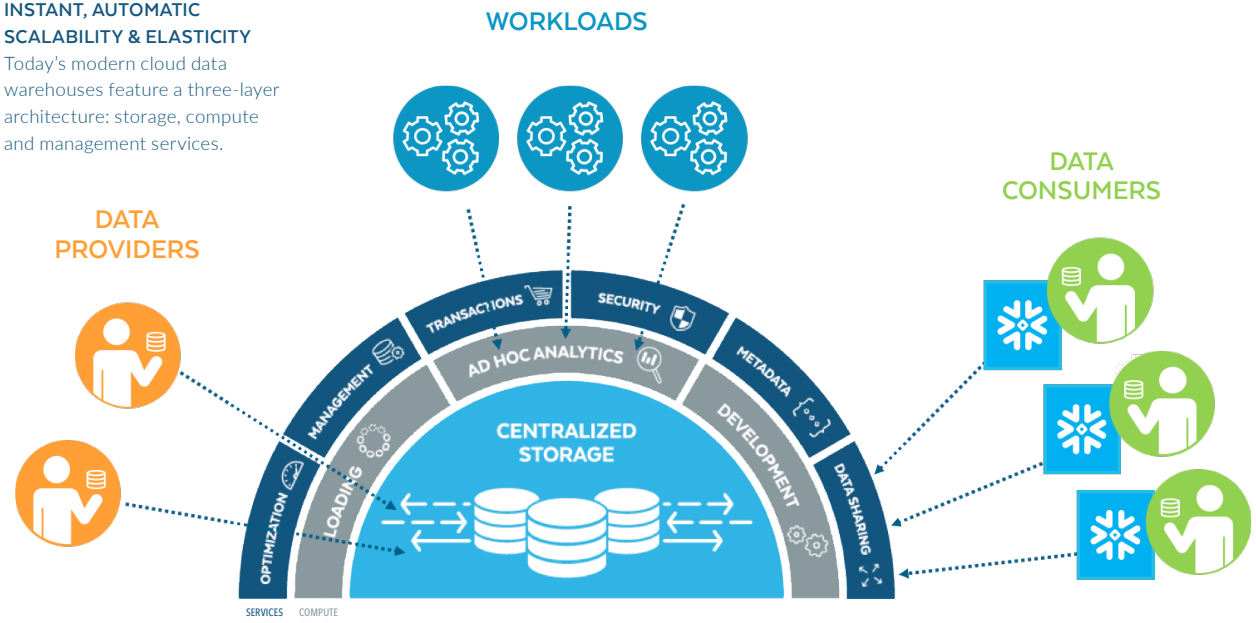
STORAGE: UNLIMITED, ELASTIC STORAGE IN THE CLOUD

The foundation of a modern data warehouse's storage layer is cloud storage. Current cloud storage services, such as Amazon Simple Storage Service (S3), offer scalability, resiliency and throughput as good or better than leading solutions for on-premises data centers, and all at a fraction of the cost. Amazon S3, for example, is designed to deliver 99.999999999% durability, and scale past trillions of objects worldwide.[12] Data warehouse architectures built for the cloud take advantage of this storage architecture by using it to hold incoming diverse data, imported tables and SQL query results.

Unlike Hadoop, whether deployed in the cloud or on premises, a modern cloud data warehouse does not co-locate storage and compute. Storage and compute resources are fully independent, yet completely integrated as a system, and can scale individually. Compute can easily scale without committing to and paying for more storage. Conversely, storage can scale up or down without disrupting compute performance and without redistributing data. As a result, data can be loaded or unloaded without impacting running queries.

INSTANT, AUTOMATIC SCALABILITY & ELASTICITY

Today's modern cloud data warehouses feature a three-layer architecture: storage, compute and management services.



[12] Source: Amazon Web Services. https://aws.amazon.com/s3/

ADDITIONAL RESOURCES

For more information about the benefits of data warehousing built for the cloud, and the difference between cloud-washed and built-for-the-cloud architectures, download the eBook, "The Data-Driven Enterprise Done Right: How to reinvent your analytics with data warehousing built for the cloud."

To learn more about how semi-structured data can be ingested into a modern cloud data warehouse, and made available for SQL queries, download the eBook, "How to Analyze JSON with SQL: Schema-on-read made easy."

Cloud data warehouse architecture, *(cont'd)*

Virtual computing and integrated services

COMPUTE

LEVERAGE CLOUD SCALE AND ELASTICITY TO DELIVER ENORMOUS PROCESSING POWER

In a modern cloud data warehouse, the compute layer is designed to process even enormous quantities of data with maximum speed and efficiency. Compute resources can be scaled up and down at any time because of cloud elasticity. But in a data warehouse built for the cloud, it's easy to create multiple independent compute clusters to handle different workloads without contention. Each compute cluster can operate on the same data simultaneously because of the separation of compute and storage. And each compute cluster retrieves the minimum data required from the storage layer to satisfy queries. As data is retrieved, it's cached locally to improve the performance of future queries.

UNLIMITED SCALE AND CONCURRENCY

This modern cloud architecture makes it possible for a data warehouse to support high levels of concurrent processing — a critical requirement for analytics and reporting. Hadoop platforms are architecturally challenged to accomplish this same level of scale and concurrency.

HOW MULTIPLE COMPUTE CLUSTERS SIMULTANEOUSLY ACCESS THE DATA IN THE STORAGE LAYER.

Multi-cluster, shared data:
Centralized, scale-out storage. Multiple, independent, compute clusters

- **Warehouse 1** can operate on a query.
- **Warehouse 2** — another, separate compute cluster — can simultaneously load or unload data without impacting the query running in warehouse 1.
- **Warehouse 3** can be created to support a separate workgroup that operates on the same data as the workgroups in warehouses 1 and 2, without affecting their query performance.

MANAGEMENT SERVICES

GLOBAL MANAGEMENT AND COORDINATION

In a data warehouse built for the cloud, the services layer is a critical part of the architecture. This layer provides global management of infrastructure, data, metadata, processing, security and availability across the environment.

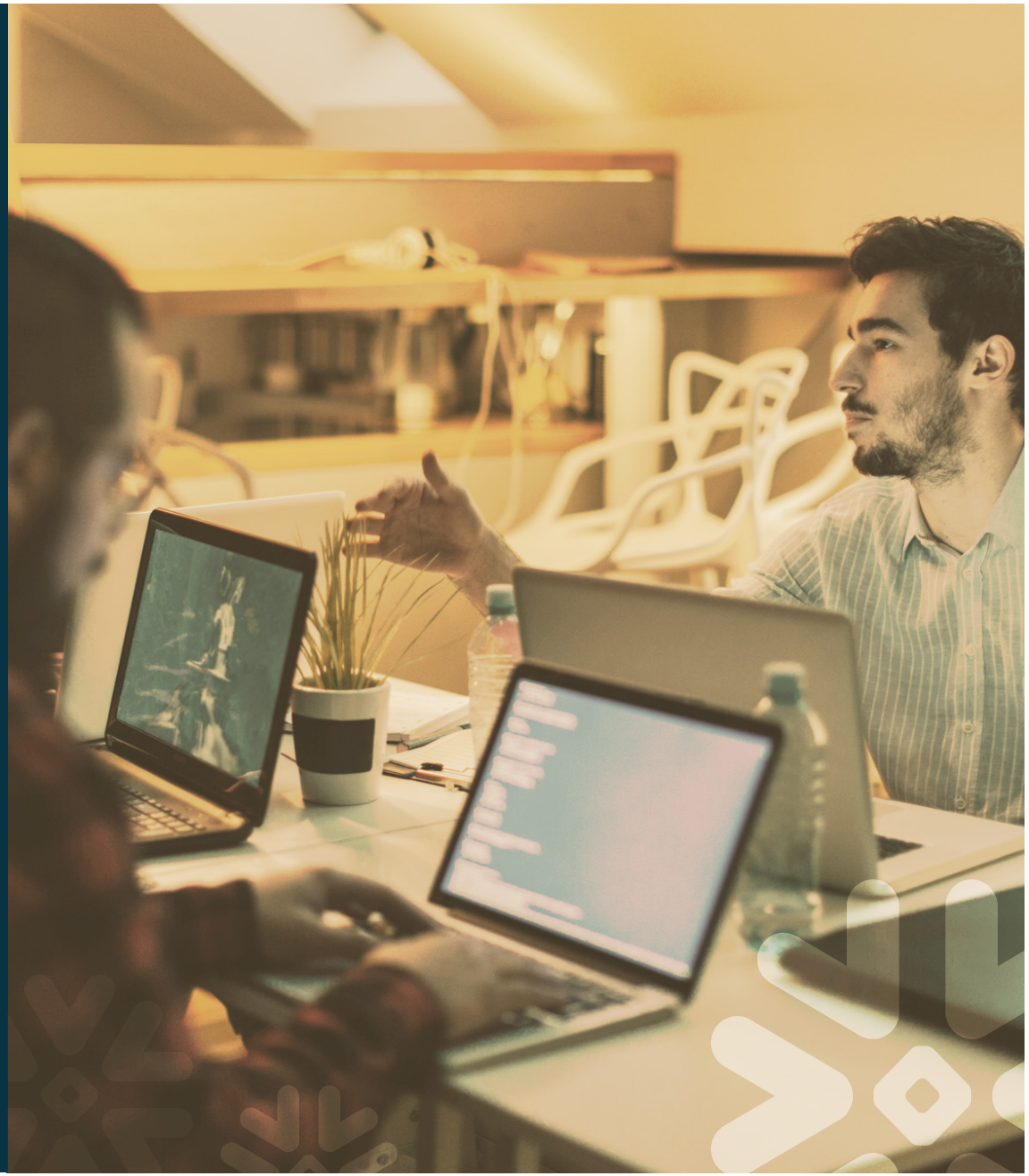
The services layer plays a key role in supporting all SQL DML[13] and DDL[14] functions. Queries are compiled within the services layer, with metadata used to determine the data that needs to be scanned.

The services layer coordinates transactions across all jobs, making it possible for multiple compute clusters to process data in a consistent, scalable way:

- ACID-compliant transactional integrity is enforced.
- Read operations (SELECT) always see a consistent view of the data, and write operations never block readers.
- Transactional integrity between compute clusters is achieved by maintaining all transaction states within the services layer.

SECURITY MANAGEMENT

The data processing services layer also manages and enforces all security features. It enforces the access control restrictions configured by customers, protecting both access to data and processing operations. It also manages data encryption, encryption keys and supports integration of customer-managed keys to ensure data security.



[13] Data Manipulation Language
[14] Data Definition Language

Reduce the administration burden

"Zero-admin" management does the work for you

The value proposition is simple: Built-for-the-cloud data warehouses are architected to automatically handle day-to-day management tasks, alleviating the administration burden. "Zero-admin" allows attention and IT resources to shift from time-consuming, low-level management tasks to deriving benefits and insights from enterprise data.

This dramatically contrasts with Hadoop's inherent complexity, which requires administrators to manage capacity planning, resource allocation, performance optimization and more. The right cloud-built data warehouse enables organizations to start small and grow to practically unlimited storage and compute resources at any time. Typical data warehouses built for the cloud can start with zero data and grow incrementally as business needs dictate. You pay only for the resources consumed.

SIMPLE VS. COMPLEX: A NIGHT-AND-DAY DIFFERENCE

The table below illustrates the differences in evaluation criteria between a data warehouse built for the cloud solution (software-as-a-service [SaaS]) and a Hadoop platform (on-premises or cloud-based).

CRITERIA	MODERN CLOUD DATA WAREHOUSE	HADOOP
DATA TYPE SUPPORT	Structured (CSV, XLS, etc.), semi-structured (JSON, Avro, XML), and Parquet data stores	Structured, semi-structured, unstructured (audio, media, text), plus Parquet and ORC data stores
OPERATIONAL ANALYTICS, DATA EXPLORATION AND DATA WAREHOUSE	Yes, all on one system	No, separate solutions required
DATA QUERY	Full ANSI SQL, DML (UPDATES, DELETES, etc.) and DDL (DROP, etc.)	Usually NoSQL, limited SQL, not full database DML or DDL
PRICING: CAPEX	No, fully OpEx	Yes, if on-premises; only OpEx if in the cloud
SCALABILITY	Very granular, compute is independent of storage; virtually unlimited	Storage and compute locked together; must pre-size
PRICE/PERFORMANCE & OPEX	Flexible and variable without taking down service; SLA-capable	Fixed; SLA's tied to cluster availability and amount of resources. For cloud, can't turn off compute pricing unless cluster is taken down.
CONCURRENCY	Virtually unlimited scaling for concurrency	Very limited concurrency
ADMINISTRATION	Zero-admin	Heavy; must provision environment
SECURITY	Encryption included	Depends on distribution components; potential performance impacts if implemented

SQL AND NOSQL WORLDS ARE STARTING TO CONVERGE

While Hadoop can support unstructured data types, the dominant use cases for big data analytics (i.e., the 80-20 rule) requires processing both structured and semi-structured data. Most big data analytics is focused on semi-structured data including weblogs, clickstreams and Internet of Things (IoT) data. Interestingly, these demands prompted Hadoop's creation but NoSQL was the chosen method early on.

The right cloud data warehouse can straddle both worlds. It can easily and natively ingest diverse structured and semi-structured data, eliminate the burdens of data transformation and provide familiar SQL-based analytics on the data, all in one solution. NoSQL is not required.

In a data warehouse built for the cloud, enterprises don't need to choose between SQL and NoSQL technologies. They can outline a very simple reference architecture that defines most data types flowing directly into the cloud data warehouse.

HANDLE SEMI-STRUCTURED DATA WITH EASE

Furthermore, semi-structured data, such as JSON files, can be ingested into a modern data warehouse built for the cloud and automatically given a compressed, columnar structure, without having to write any special scripts. In addition, this semi-structured data can be easily accessed via SQL queries. These technology breakthroughs present major improvements over labor-intensive data preparation in both Hadoop and traditional data warehouses.

Getting started is easy

A simplified solution with cloud data warehousing

In the past, SQL champions were faced with a painful "moment of truth" when setting up a warehouse to analyze web or IoT data. Did the IT team have the skills and knowledge to spin up, operate and maintain a Hadoop cluster? Could they write Java scripts? Today, SQL pros can simply go to the website of a modern cloud data warehouse vendor and sign up for the service and use it immediately. Table 3 summarizes the necessary steps for common usage scenarios.

ALL-NEW DATA WAREHOUSE

Use case: A new data project with your company or for a new company building out its data architecture

- Initiate service and activate it when ready.
- Set-up/define data type parameters.
- Select and configure a data loading service. No transformation required.
- Point the data loader to the cloud data warehouse.
- Start ingesting data.
- Once the data is committed, SELECT the data, issue the query and start analyzing.

AN EXISTING DATA WAREHOUSE

Use case: If you already have a data warehouse, decide whether you will import/migrate all your data or a portion of it.

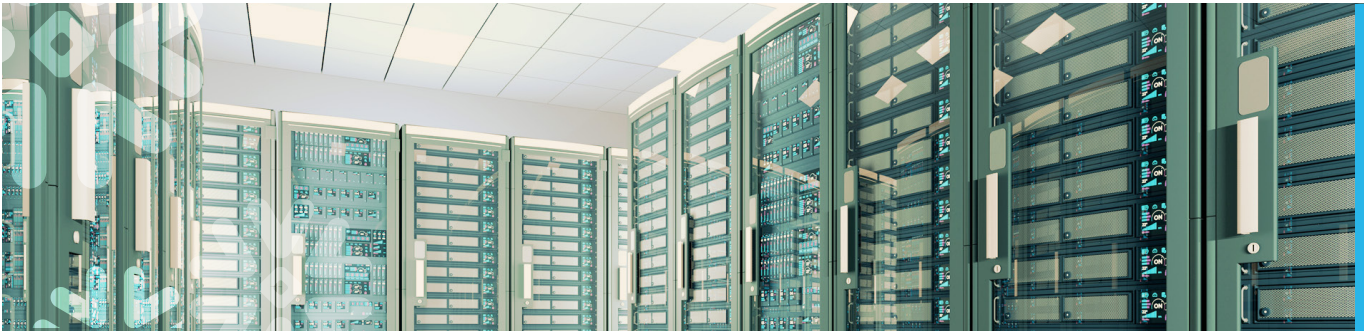
- Follow all the same steps as for a new data warehouse
- As needed, enlist a qualified system Integrator (SI) for migration assistance.

REAL-TIME TRANSACTION DATA PROCESSING OR STREAMING ANALYTICS

Use case: When real-time transaction processing is required for semi-structured data

- Add a Spark or NoSQL front end for the transactional and streaming portion.
- Move the data to the cloud for warehousing after the real-time processing is executed.

The benefits of data warehousing built for the cloud



A summary of all the beneifts

Compared with implementing and managing Hadoop, a traditional on-premises data warehouse or one that's been "cloud-washed," a data warehouse built for the cloud can deliver a multitude of unique benefits. In addition to eliminating the unintended consequences, support capacity issues, legal liability and security vulnerability of Hadoop, a modern cloud warehouse offers:

- **Fast deployment:** A cloud-built data warehouse can go live in just minutes. Deploying Hadoop requires extensive planning, software customization, Java programming and server provisioning.
- **No software upgrades to manage:** Enterprise IT staff must manage Hadoop environments, and their myriad of components, creating a significant ongoing maintenance burden. On-premises and cloud-washed commercial warehouse solutions take a standard "waterfall" development approach to functionality updates. To enable the annual or biannual update, IT staff must take the system down or place it in maintenance mode, losing time and money. To avoid this, IT may anchor to a specific version of the software, which creates another set of headaches.

  With a modern cloud data warehouse, upgrades originate from an agile DevOps approach — incremental updates every month that avoid any disruption to customers.
- **Resources focused on what matters most:** The number of people who maintain Hadoop environments and traditional data warehouses come at an enormous expense. A zero-admin data warehouse built for the cloud can eliminate unnecessary management tasks, allowing IT staff to focus on high-impact data challenges instead.
- **Pay only for what you use:** Both on-premises and cloud-washed Hadoop and data warehouse solutions force enterprises to buy enough storage space and compute horsepower to handle demand on the busiest day of the year. What about the other 364 days? The right cloud data warehouse will require that you pay only for what you use, when you use it. Additionally, the cost of the actual storage and compute resources should be significantly lower with a cloud solution thanks to the cloud's economies of scale.

Find out more

Find out more about how data warehousing built for the cloud can take your enterprise data analytics operation beyond Hadoop to a higher level of productivity at lower cost. Find out more at **www.snowflake.net**

About Snowflake

Snowflake is the only data warehouse built for the cloud. Snowflake delivers the performance, concurrency and simplicity needed to store, analyze and share all data available to an organization in one location. Snowflake's technology combines the power of data warehousing, the flexibility of big data platforms, the elasticity of the cloud and live data sharing at a fraction of the cost of traditional solutions. Snowflake: Your data, no limits.

snowflake