Making Data Warehousing Simple

Building a Just-in-time Data Warehouse Platform with Databricks

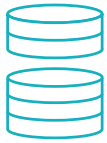
A New Approach to Data Warehousing

Enterprises are dealing with a wider variety of data at a much larger scale today, coming in at ever faster rates. In addition to the data found in their existing database management systems (DBMS) and enterprise data warehouses (EDW), enterprises need to make sense of the massive volume of semi-structured data such as sensor data, clickstreams, and logs residing in other data stores. For traditional data warehousing, this requires data teams to constantly build multiple costly and time-consuming extract - transform - load (ETL) pipelines to ultimately derive business insights.

Databricks provides a fast, simple, and scalable way to augment your existing data warehousing strategy by combining pluggable support for common data sources and the ability to dynamically scale nodes and clusters on-demand. Additionally, Databricks has built-in SSD caching to complement Apache® Spark's™ native in-memory caching to provide optimal flexibility and performance. This enables organizations to read data on-the-fly from the original data source and perform "just-in-time" queries on data wherever it resides rather than investing in complicated and costly ETL pipelines.

databricks™

Just-in-Time Data Warehouse



Elastic Scalability

- Distributed clusters on-demand
- Scale storage and compute independently
- Efficient SSD and RAM caching

Faster Time-to-Insight

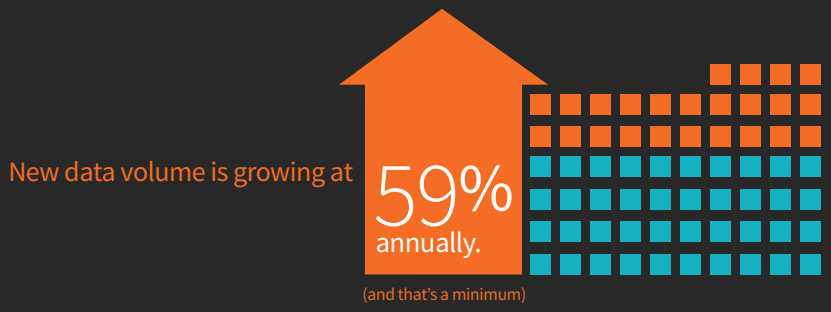
- Direct data access with schema-on reads
- Efficient support for a variety of data sources
- Flexible processing of unstructured, semi-structured, and structured data

Advanced Analytics

- Support for Python, R, Scala, and SQL
- Machine Learning, streaming, and graph processing
- Interactive notebooks for data exploration

Traditional Data Warehousing Pain Points

In today's big data landscape where a variety of data comes from a multitude of highly disparate data sources, it is fast becoming a challenge to glean timely business insights. The rigidity of traditional enterprise data warehouses only adds to the pain as companies are forced to fully architect their data warehouse in advance regardless of data volume fluctuations, making it very costly and time-consuming to maintain over time. This is causing enterprises to shift away from relying completely on traditional data warehouses to handle their big data needs.



Source: Gartner Group, Pattern-Based Strategy: Getting Value from Big Data

Further adding complexity and inefficiency is the need for data to go through an ETL process before being queried, which can be very time consuming and hamper a team's ability to query current data in a timely fashion. In many cases, enterprises ETL their data just once every 24 hours. Worse yet, as the number of data sources and the volume of the data increases, the ETL time also increases, negatively impacting when an enterprise can derive value from the data.

Source: Deloitte Consulting, The Future of Data Warehouses in the Age of Big Data

Traditional Data Warehousing Pain Points

Here is an overview of the common pain points that data and ETL teams experience when working with traditional data warehousing solutions:

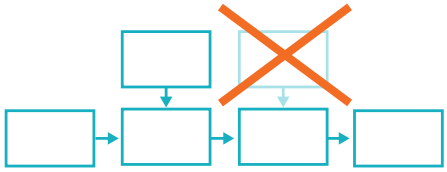
- **Inelasticity of compute and storage resources**
- **Rigid architecture that's difficult to change**
- **Limited advanced analytics capabilities**

Inelasticity of compute and storage resources

The bursty nature of traditional data warehousing (e.g. increase in query loads during fiscal month end, idle loads during holidays, etc.) requires that capacity planning to account for maximum load. The fixed size of most data warehouses results in compute and storage to scale linearly together, yet these problems are orthogonal to each other (e.g. burst in queries does not always translate to an increase in storage). This inelasticity results in an expensive conundrum: if your data warehouse is successful you cannot easily expand; if there is overcapacity, it means there are idle resources.

Rigid architecture that's difficult to change

Traditional data warehouses are schema-on-write requiring schemas, partitions, and indexes to be pre-built prior to the first query. These requirements for rigidity and for ETL pipelines to manipulate this data is not only time consuming, but costly as this limits your ability to gain access to many new data sources and datasets. With over 90% of the world's data generated in the last two years, and most of it semi-structured, you will have to expend an increasing amount of finite resources to continually augment your ETL pipelines. Additionally, you'll need to go through this process each time your data changes and your data sources are augmented.



Limited advanced analytics capabilities

Enterprises want more than what business intelligence and data warehousing provides today. More than just counts, aggregates, and trends; they need to to detect hidden patterns, forecast using machine learning, and segment customers based on features using learning algorithms. While traditional data warehouses are optimized for querying and processing against structured data, they lack the fundamental components to go beyond SQL.

How Databricks Helps You Easily Build a Just-in-Time Data Warehousing Platform

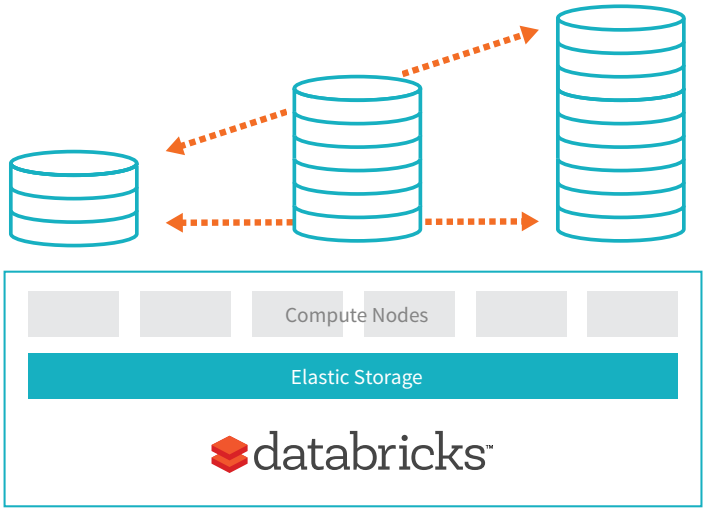
Databricks allows data teams to rationalize their data warehousing strategy by offering a just-in-time data warehouse solution designed to easily handle a wider variety of data from different sources, while enabling faster turnaround times for data analysis projects with processing speeds up to 100x faster than MapReduce at scale.

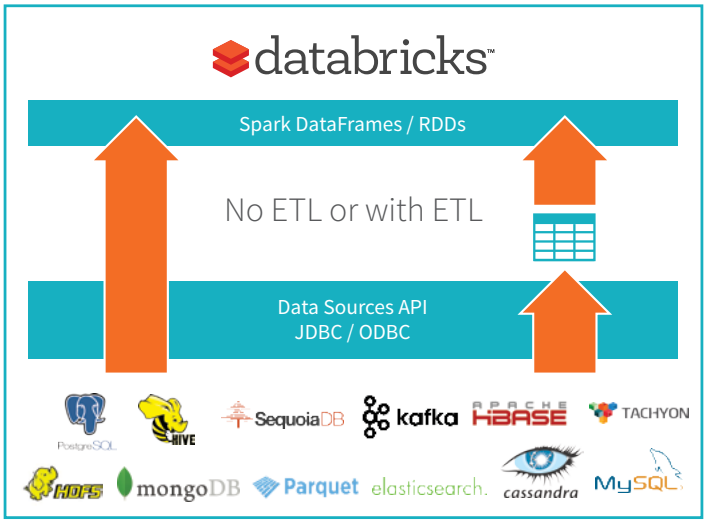
Here are examples of how an enterprise can benefit from using Databricks as a just-in-time data warehouse:

- **Scale resources on-demand**
- **Direct access to data sources**
- **Leverage advanced analytics capabilities**

Scale resources on-demand

Having on-demand clusters to scale up and down based on your query loads is a cost effective and efficient solution for bursty data warehousing queries. Databricks can provide on-demand clusters because it separates your persistent storage from your compute clusters. This not only allows you to scale your compute and storage independently, but also easily setup multiple clusters to access the same data sources. Databricks' Spark in-memory clusters have built-in SSD caching that cache your files upon extraction to minimize pre-processing and speed up your queries.





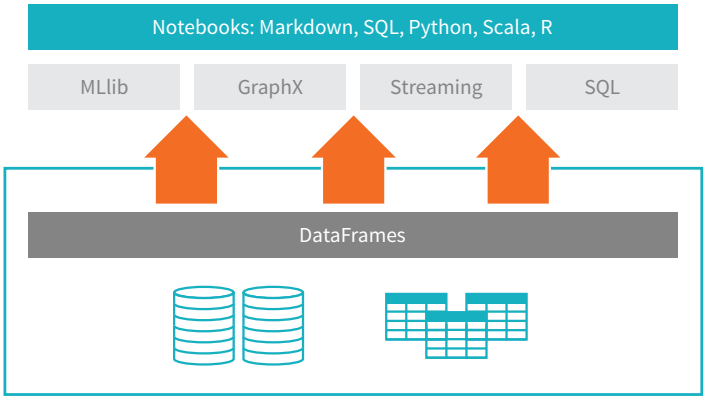
Direct access to data sources

Databricks allows you to circumvent the schema-on-write ETL process to directly access the data (i.e. schema-on-read). While you can still perform traditional ETL process, you are no longer required to perform all of this data movement. Databricks combines together Spark's Data Sources API, JDBC/ODBC connectors, and pre-installs hundreds of spark-packages to access a variety of data sources to provide you with the ultimate data access platform. With Databricks, in many cases filtering and column pruning is pushed all the way down to the data source.

How Databricks Helps You Easily Build a Just-in-Time Data Warehousing Platform

Leverage advanced analytics capabilities

The Spark ecosystem can help you solve many different data problems including machine learning (via MLlib), graph computations (via GraphX), streaming (e.g. real-time calculations), and real-time interactive query processing with Spark SQL and DataFrames. The ability to use the same framework to solve many different problems and use cases allows data professionals to focus on solving their data problems instead of learning and maintaining a different tool for each use case. Databricks enhances the Spark ecosystem by providing native and third party visualizations within notebooks that support multiple languages. The interactive notebooks allows for your data teams to work collaboratively (instead of in silos) to deliver timely insights.



To learn more about how Databricks can help you build your just-in-time data warehousing solution, please take a look at the following materials which provide an overview of the Databricks platform and how it can help data teams focus on getting value out of their data.

Simplifying Spark Operations with Databricks
Databricks Primer
Databricks Feature Primer
Databricks Security Primer

Just-in-time Data Warehousing Scenarios

With Databricks, enterprises can utilize more of the data they already have and accelerate the turnaround time for their big data analytics projects. Here are examples of companies who have leveraged Databricks to solve their biggest data warehousing challenges.

Elastic Infrastructure

There are significant cost-benefit savings when you only have to pay for what you use. For example, it is common to see peak query loads during fiscal period ending for financial reconciliation. In traditional data warehousing scenarios, you will need to build out infrastructure to support the maximum processing and query load which is costly.

With Databricks, you can build a just-in-time data warehouse that can be built up, replicated, and torn down whenever you need it. For example, during fiscal period end, you can spin up extra nodes or clusters to allow for the influx of query requests. Once the reconciliation is done, you can simply spin down or tear down the nodes and clusters no longer needed.

Just as important, instead of having dedicated DevOps personnel to maintain your big data infrastructure, Databricks offers a simple management UI to instantly create, scale up, and teardown Spark clusters as needed.

This is how MyFitnessPal was able to implement a new feature called “Verified Foods” to provide more accurate nutrition information, an order of magnitude speed improvement, and have improved team efficiency and productivity. For more information, please refer to the Databricks Customer Case Study: MyFitnessPal.

Beyond Data Warehousing

Utilize your team’s data warehousing expertise as a starting point for your data platform. This allows your data warehousing specialists, ETL engineers, and analysts to have quick and easy access to their data. Meanwhile, your data scientists can go beyond data warehousing and branch into advanced analytics and streaming. A great example is Radius Intelligence and how they imported machine learning algorithms for testing and data quality verification as well as used GraphX to better visualize and understand changes to Radius’ business listing index. For more information, please refer to the Databricks Customer Case Study: Radius Intelligence.

Just-in-time Data Warehousing Scenarios

More Effective Collaboration

Traditional data warehousing techniques and team structures often result in siloed and disparate environments. Yet, building data products effectively requires a collaborative engineering environment. With the adoption of Databricks, Celtra has enabled teams from Engineering, Product Management, and QA to perform complex data analysis collaboratively via interactive notebooks, leveraging their massive production data to improve product design, address anomalies rapidly, and fine-tune the performance of production systems. For more information, please refer to the Databricks Customer Case Study: Celtra.

Accelerated Time-to-Market

The ability to deploy and scale a data warehouse on-demand allows your team to focus on your data problems rather than the infrastructure. For example, Yesware built its production data pipeline in just under three weeks by building it with Databricks. Their prior solution took 12 hours to process 90 days of data; with Databricks they are processing over 180 days of historical data in two hours. For more information, please refer to the Databricks Customer Case Study: Yesware.

Immediate Results and Business Insights

To explore, visualize, or analyze results from a traditional data warehousing solution, you often need to custom fit disparate solutions together. For example, it's often necessary to have separate tools for reporting, ad-hoc analysis, and dashboarding. The ability to visualize, iterate, and collaborate with multiple disparate software tools is a difficult endeavor resulting in data teams focusing on the tools instead of their data. But with Databricks, customers are able to effectively collaborate and visualize their results immediately. For more information, please refer to the Databricks Customer Case Study: Sharethrough.

The Just-in-Time Data Warehouse. Delivered.

For enterprises who are experiencing the pains of traditional data warehouses, Databricks provides data teams with the ability to scale storage and compute clusters to meet demand and speed time-to-insight through flexible support for multiple data formats and sources. By implementing Databricks as a just-in-time data warehouse, teams will eliminate productivity bottlenecks, enabling them to access the data when they need it for fast analysis, and to simplify the management of clusters for faster time-to-market.

Get started with Databricks for just-in-time data warehousing today with a free trial.

databricks™

making big data simple