



aws SUMMIT

INDIA | MAY 25, 2023

A N A 0 0 3

Real-time analytics using Apache Druid as a Service on AWS

Abhishek Agarwal
Senior Engineering Manager
Imply



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

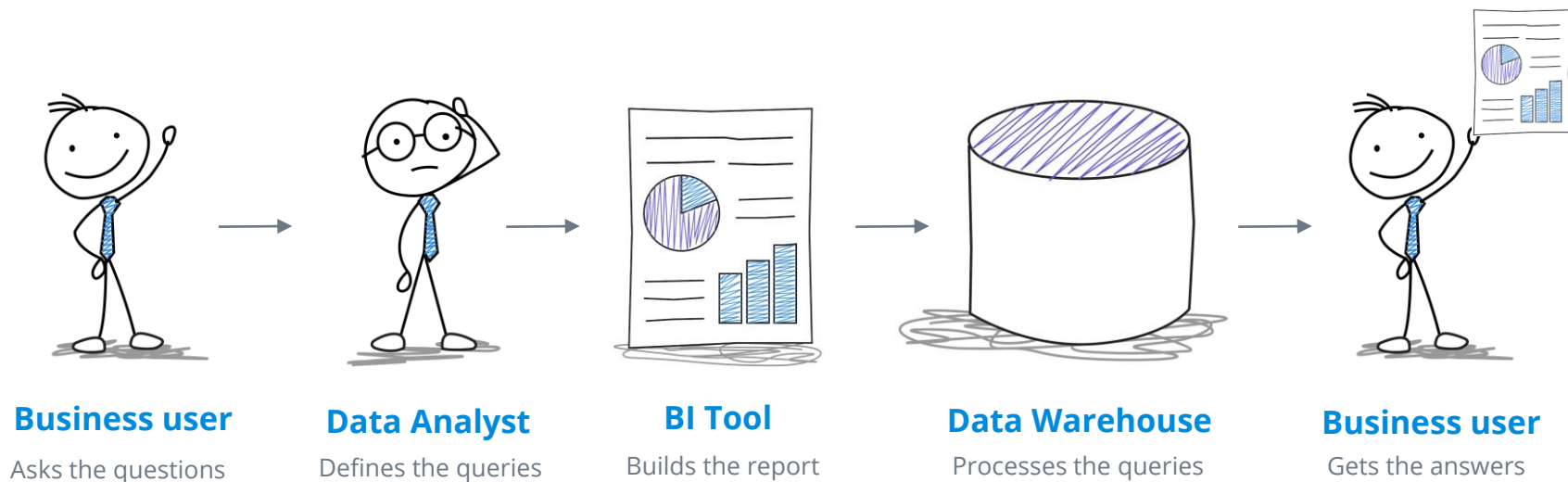


The Database for Modern
Analytics Applications

Real-Time Analytics at Scale with Apache Druid

Analytics are expanding
beyond reports and BI

The classic analytics reporting workflow



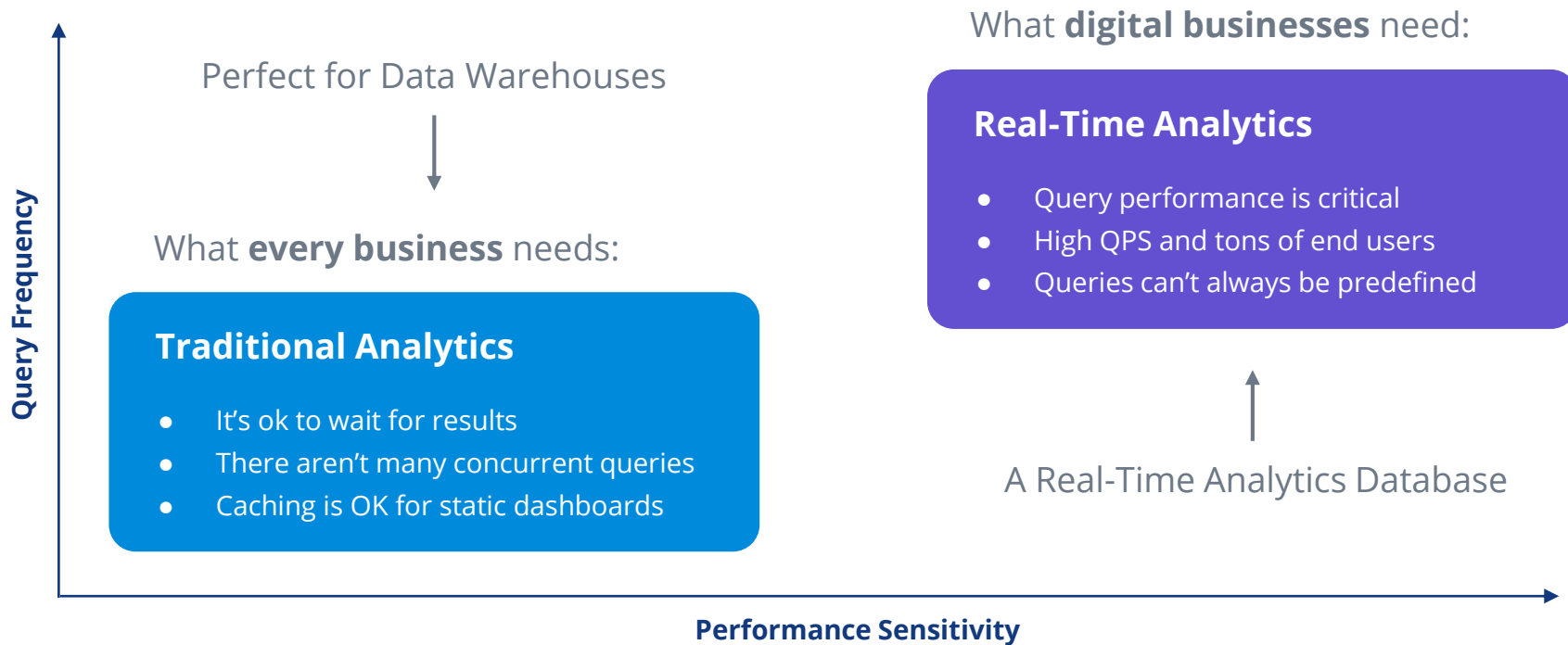
Fundamental Assumptions

It's OK to wait for the results.

There aren't many concurrent queries.

Caching is OK for static dashboards.

But analytics are expanding beyond reports



Supporting
150,000
Global Customers

Edge Intelligence Observability Platform

salesforce



Product Owners

Trillions of
rows ingested
every day



Engineers

80,000
concurrent
users



Customer Service

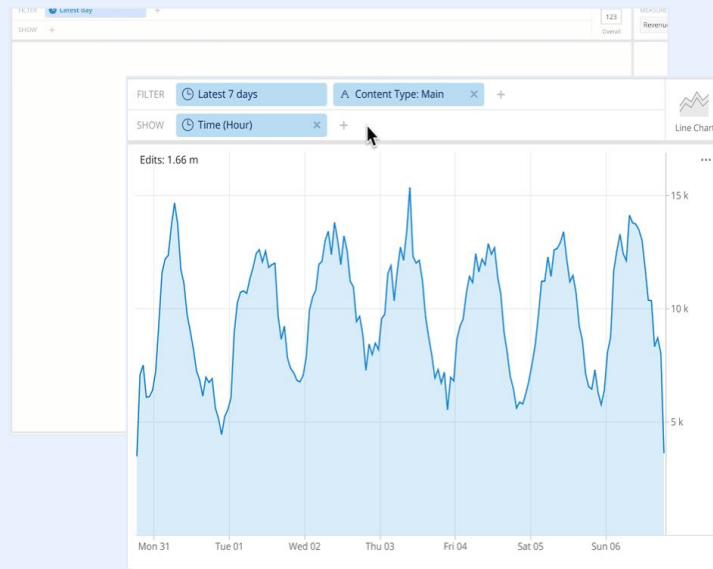
Query Results
within
seconds

Fundamental Design Principles



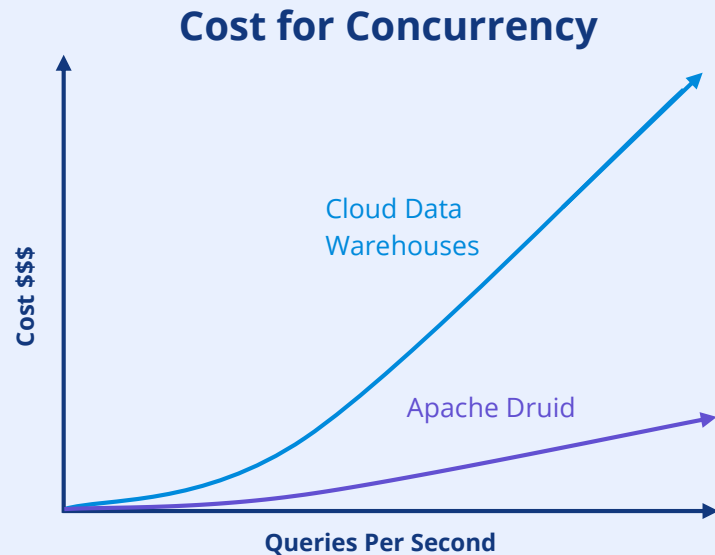
Sub-second Response at Any Scale

Interactive analytics on TB-PBs of data



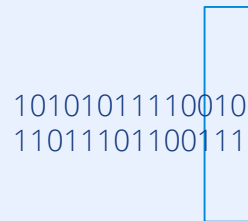
High concurrency at the lowest cost

100s to 1000s QPS via a highly efficient engine



Real-time and historical insights

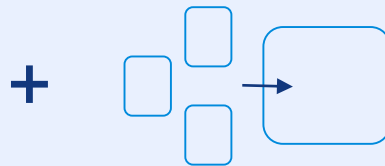
True stream ingestion for Kafka and Kinesis



10101011110010
11011101100111

The diagram shows two lines of binary code (10101011110010 and 11011101100111) positioned to the left of a vertical rectangular box, representing a continuous stream of data being ingested.

Stream Ingestion



Batch Ingestion

Examples of real-time analytics applications

Operational
Visibility at Scale



ICE Security Ops Platform

Customer-facing
Analytics



Citrix Analytics Service

Unrestricted
Data Exploration



Salesforce Edge Intelligence

Real-time
Decisioning



Reddit Real-Time Ads

Powered by  druid

Developers absolutely love Druid...

How Netflix uses Druid for Real-time Insights to Ensure a High-Quality Experience



Netflix Technology Blog

Follow



Mar 3, 2020 · 9 min read



CONFLUENT

NOVEMBER 8, 2021



ZOHREH KARIMI



HARINI RAJENDRAN

Why Apache Druid?

Before migrating to Druid, we used a non-time-series NoSQL database to store and query our telemetry metrics. As the volume of data grew, our legacy pipeline struggled to keep up with our data ingestion and query loads. It also wasn't suitable for some new use cases like the Confluent Cloud Metrics API we have on top of Druid today that customers query directly.



Dun Lu

Oct 22, 2020 · 10 min read

Delivering High-Quality Insights Interactively Using Apache Druid at Salesforce

Performing OLAP (Online Analytical Processing) data analysis over an ever-growing data set might not seem as challenging as launching a rocket nowadays, but delivering high-quality insights at a large scale is never a trivial job. As the Edge Intelligence team in Salesforce, our goals are to:

Powering real-time data analytics with Druid at Twitter

By Ruchin Kabra and Chunxu Tang

Druid at Twitter

An important characteristic of Twitter is its real-time nature. Consequently, many of Twitter's projects need real-time analytics as a platform service. In recent years, Twitter's data platform team has evolved **Druid** as a real-time centralized analytics platform at Twitter. Druid is a real-time analytics database designed for fast slice and dice analytics on large datasets. It is most often used as a database for powering use cases where real-time ingestion, fast query performance, and high uptime are important.

Trusted technology with a vibrant community

14,000+

Community Members

500+

Active Contributors

1,900+

Companies using Druid

150%

YoY Increase in Community Activity

Imply: The complete experience for Apache Druid



Commercial Distribution

Management, monitoring,
and early features and
patches



DBaaS, Hybrid or Software

Imply Polaris and
hybrid-managed service



Committer- Driven Expertise

24/7 support with 100% of
the original Druid creators

Plus, Imply Pivot to accelerate application development

With Imply, devs get rapid time to value and success with Druid



Imply Polaris

The Cloud Database Service
for Apache Druid

————— And for OS Druid Users —————



Most
Affordable



Most
Secure



Best Time
to Value

One UI for a single development experience

Dev/Test

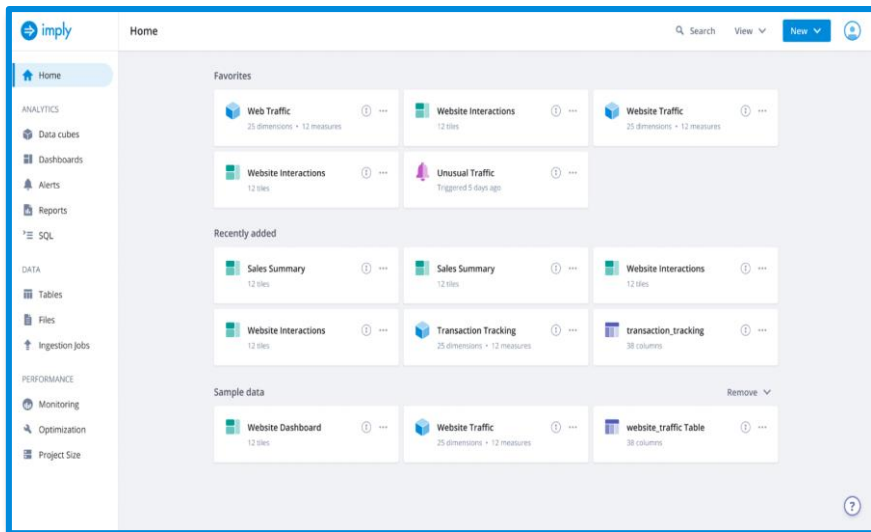
Write queries with instant responses

Visualization

Create interactive experiences w/ a suite of fully customizable visualizations

Scale

Effortless scale your project with zero downtime



Ingestion

Built-in push ingestion sends data from anywhere with no setup

Usage

Know what you are spending with consumption visibility for full transparency

Monitor

Curated dashboards to monitor your projects performance

ImPLY Polaris is a fully managed cloud service



No Infrastructure Management

No need to talk to DevOps or IT or procure any infrastructure



Ready In Seconds

Sign up and you are ready to start loading data and extracting insights



ImPLY Polaris



Scale Up and Down

Effortlessly scale your project with zero downtime



Continuous Improvements

Get the latest features and security patches with no disruption

Flexible Data Ingestion Options

Stream Ingestion



Native, pull-based ingestion capability from existing Confluent Cloud deployments



Push events to Polaris directly from your own app using the Events API



HTTP-based Kafka Connector which uses the Polaris Push API endpoint



Amazon Kinesis

Native, ingestion stream by reading data from a Polaris-Kinesis Connection

Batch Ingestion



Native batch ingestion from Amazon S3

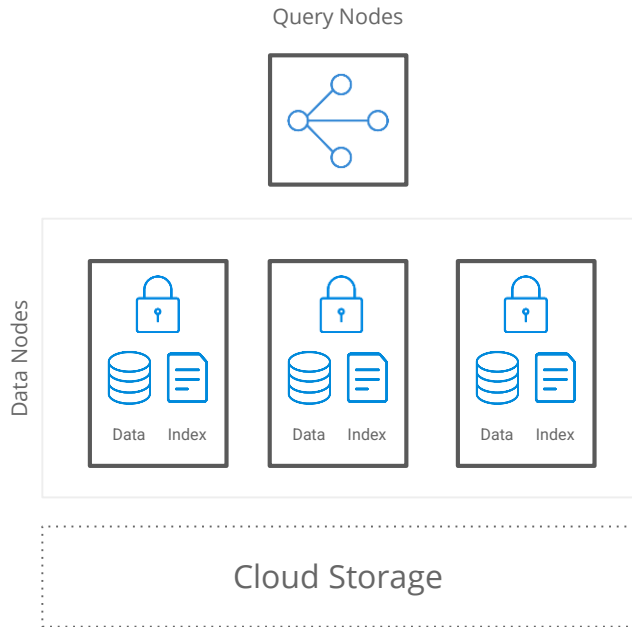


Load data from one Polaris table into another table



Upload files to the Polaris staging area & ingest from uploaded files
JSON | PARQUET | ORC
Avro | CSV | TSV

Developers get built-in protection and security



Automatic recovery

Nodes are automatically replaced and balanced upon failure. Data are automatically reload from cloud storage [S3].

Secure by default

All data are encrypted by default (in flight/at rest). SOC 2 Type 1 and HIPAA certification. Role-based access control to manage authorization

Built-in protection & DR

Replication & backup come automated with no set up required. Data are stored in cloud object storage [S3], easily recoverable upon disaster.

Online upgrades

Take advantage of the latest features and always be on the current version with no disruptions.

Thank you!

Abhishek Agarwal

Senior Engineering Manager

Imply



Please complete the
session survey

