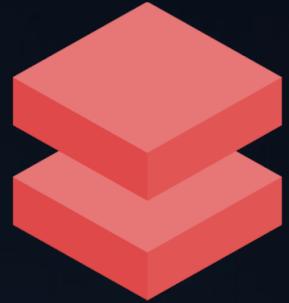




ONE WAY
SOLUTION



One Way Solution **TDW, MDW & Lakehouse**

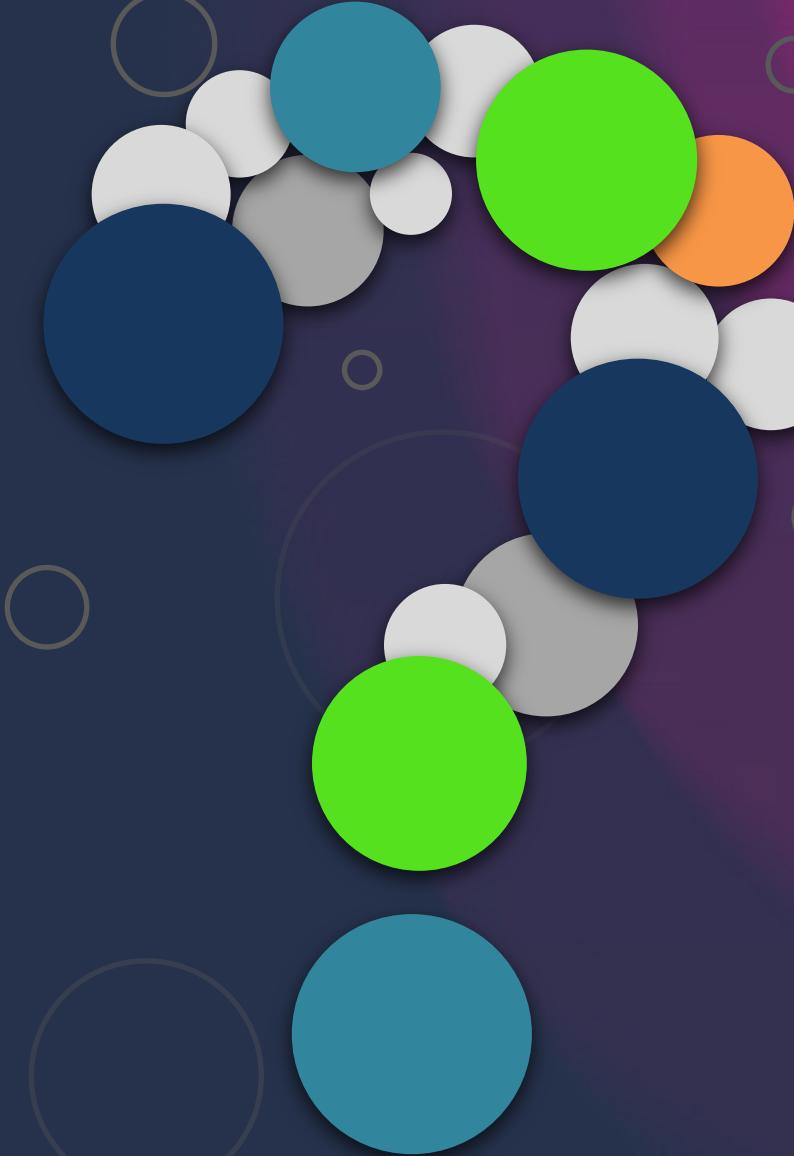
Data Engineering – [Day 4]

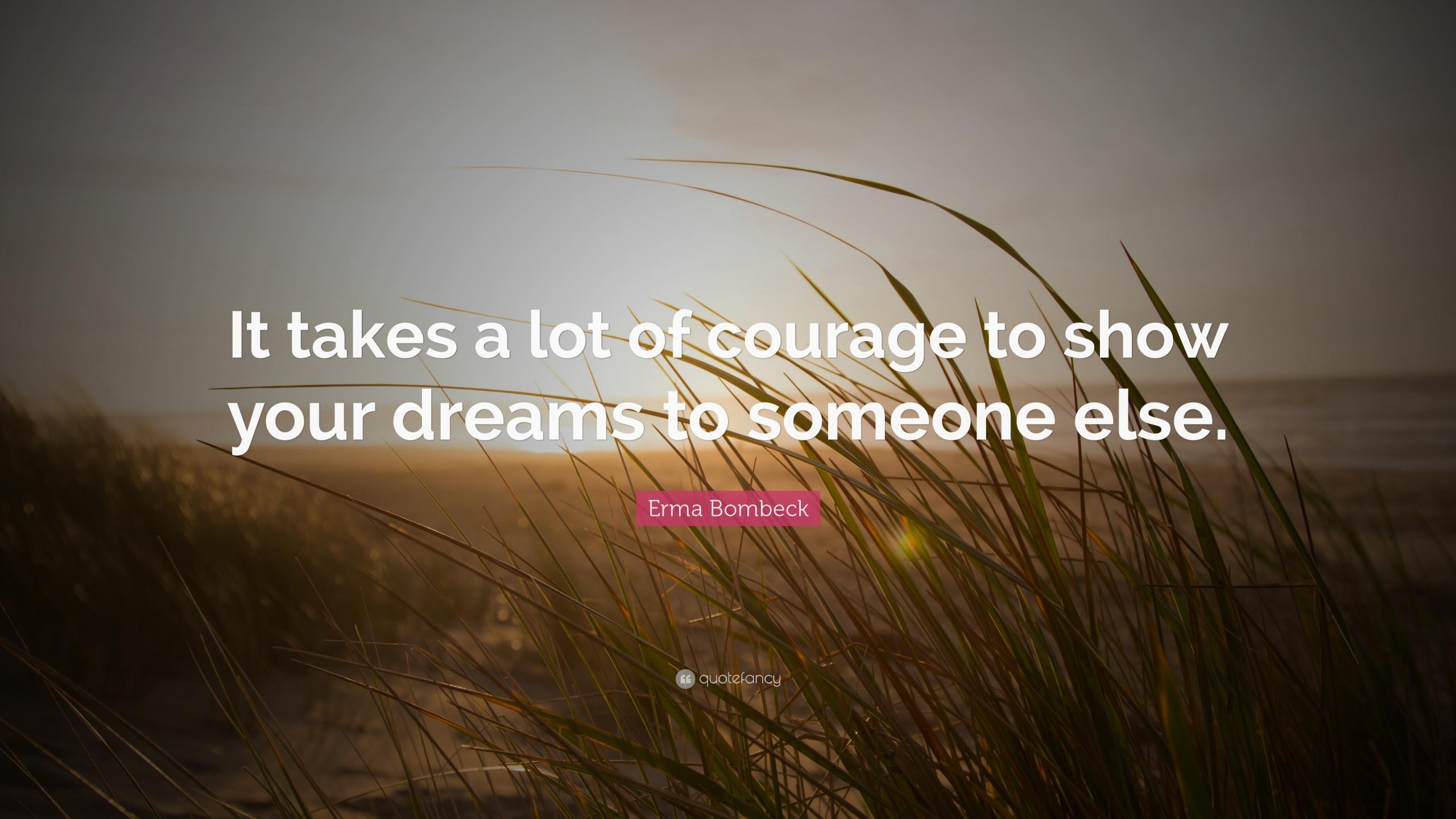


LUAN MORENO

CEO & CDO

Data Engineer & Data Platform MVP



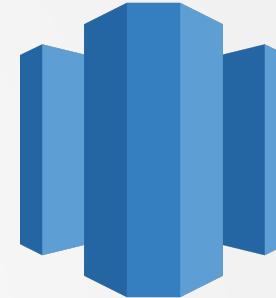


**It takes a lot of courage to show
your dreams to someone else.**

Erma Bombeck

ETL vs. ELT

Extract/Transform/Load (ETL) is an integration approach that pulls information from remote sources, transforms it into defined formats and styles, then loads it into databases, data sources, or **Data Warehouses**.



Extract/Load/Transform (ELT) similarly extracts data from one or multiple remote sources, but then loads it into the target **Data Lake** without any other formatting. The transformation of data, in an ELT process, happens within the target database. ELT asks less of remote sources, requiring only their raw and unprepared data.



Traditional Data Warehouse [Dw]

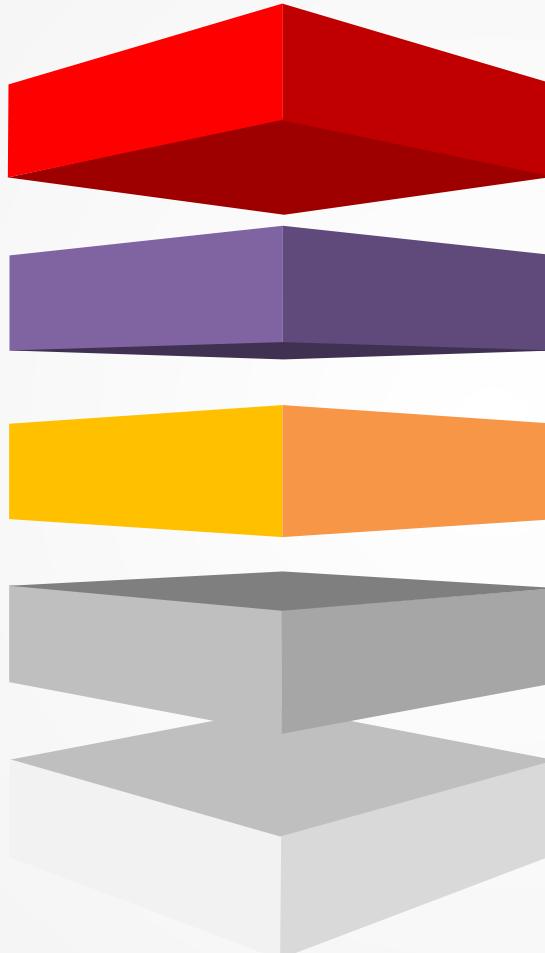


History

- 1980s from IBM Researchers
- Operational System for Decision Support
- Bill Inmon

Data Storage & Retention

- Store Current & Historical Data
- GB to TB of Data in a Single Place



Design Methods

Bottom-Up = Data Mart [Ralph Kimball]
Top-Down = Data Warehouse [Bill Inmon]

EDW [Enterprise Data Warehouse]

- Used for Reporting & Data Analysis
- Component of Business Intelligence Solution
- Central Repository from Multiple Data Sources

Techniques

- ETL [Extract, Transform & Load]
- Apply Business Logic
- Use Stage Area [Stage]
- Use [ODS] to Keep Relational Structure

Dimensional Modeling

- Business Process
- Grain
- Dimensions
- Facts
- Star & Snowflake Schema

Star Schema Model



Model

- Separates Business Process Data into **Facts**
- **Dimensions** with Descriptive Attributes
- Measurable & Quantitative Data



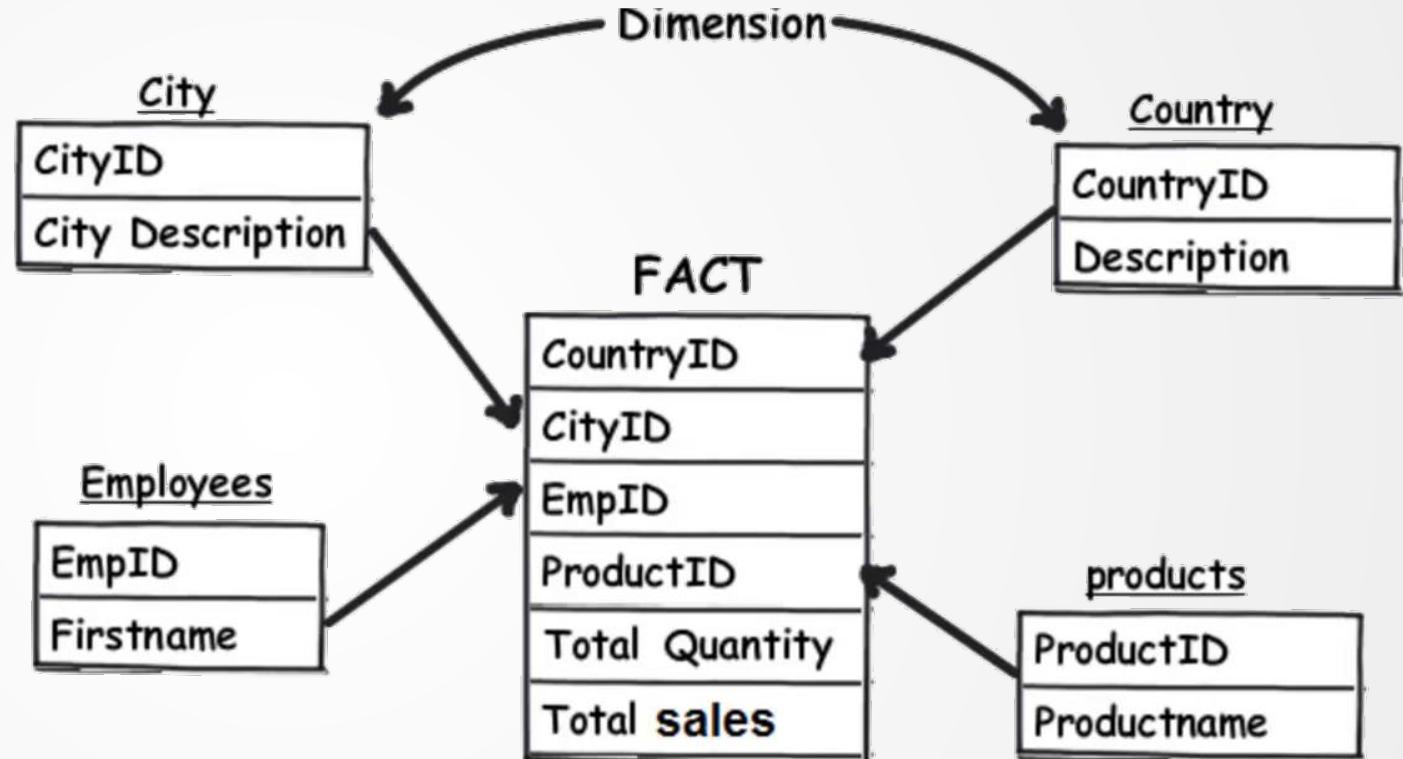
Benefits

- Simpler Queries
- Simplified Business Reporting Logic
- Query Performance Gains
- Fast Aggregations
- Feeding Cubes [OLAP]



Disadvantages

- Data Integrity [De-Normalized State]
- Batch Processing Load Fashion



Azure SQL DB



Cloud Database as a Service



Intelligent Relational Cloud DB

- Managed by Microsoft Azure
- Intelligent Features for Performance & Administration



Fully Managed

- General Purpose Database
- Global Scalability
- Near-Zero Administration
- Dynamic Scalability with Zero Downtime



Cloud-First Strategy



- Microsoft SQL Server Code Base
- New Features



Features

- Columnar Storage [[ColumnStore](#)]
- In-Memory Technologies
- Data Sync
- Multi-Model Capabilities
- Job Automation
- Transactional Replication
- Temporal Tables
- HyperScale [100 TB]



Azure SQL DB as a Traditional Data Warehouse [TDW]

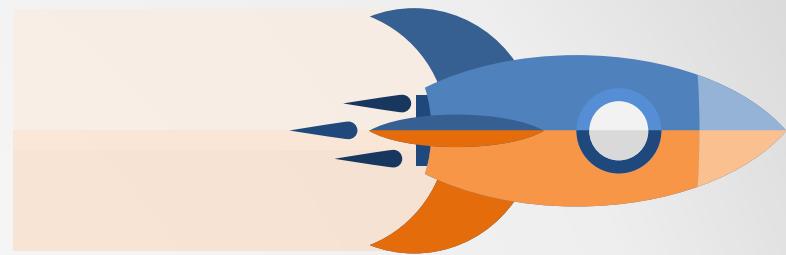


Doubt kills more dreams
than failure ever will.

Suzy Kassem

Data Warehouse [2.0] – Modern Data Warehouse [Dw]

A modern data warehouse lets you bring together all your data at any scale easily, and to get insights through analytical dashboards, operational reports, or advanced analytics for all your users

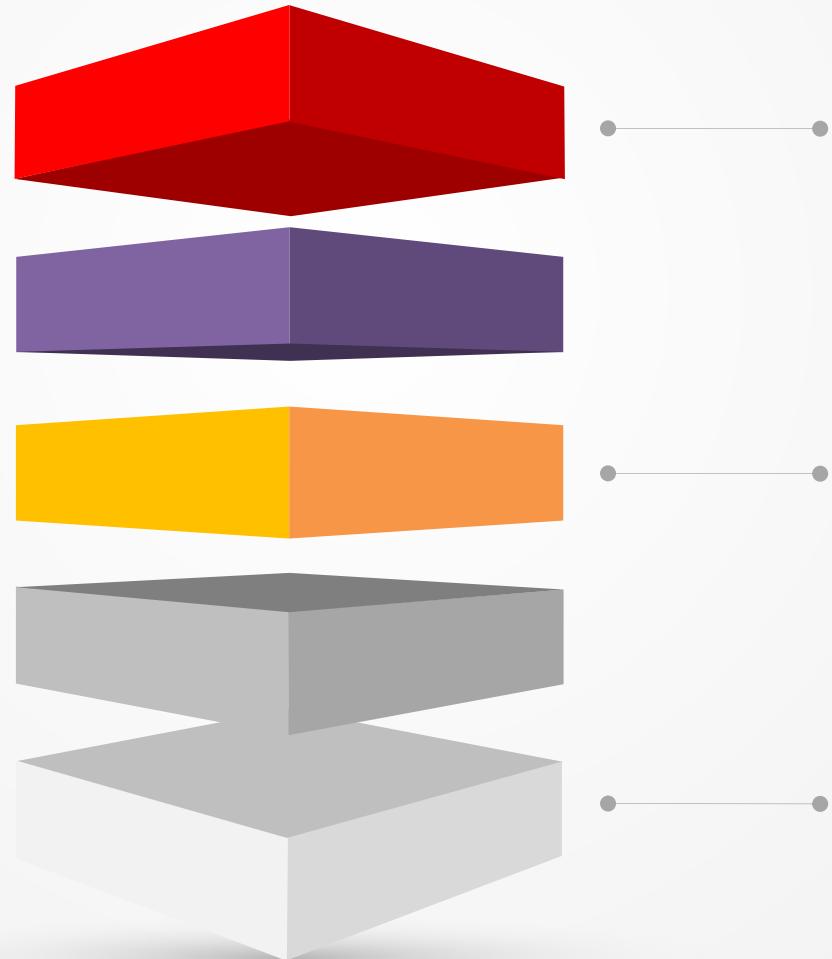


- Analytics Platform for Enterprises
- Scalability – Horizontally vs. Vertically
- PaaS & SaaS
- SQL-Like Interface [SQL]



Physical Hardware

- Massively Parallel Processing [MPP]
- “Loosely Coupled” & “Shared Nothing”



Columnar Data Storage Type

- Batch-Processing Mode
- Compression Benefits
- I/O Reduction Operations
- Index Bitmap

Jack
Wu
Sam
Jen
20
32
45
17
Montreal
Winnipeg
Toronto
Vancouver
22000
35000
43000
22000

Cloud-Based Data Warehouses Systems



- Amazon Redshift
- Azure Synapse Analytics
- Google BigQuery
- Snowflake



Elastic Compute & Storage

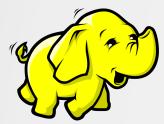
- [PaaS] – Platform-as-a-Services
- Distributed Computation Architecture
- Computation <> Storage



Caching

- Sub-Second Response Time
- Performance Boost

Use-Case for Microsoft Azure

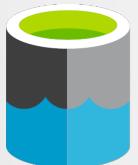
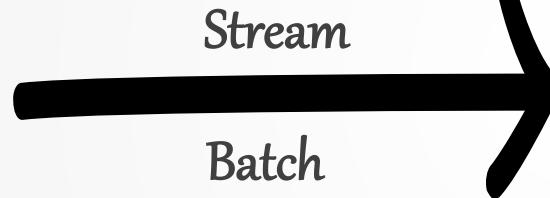


HDInsight



Easy & Cost-Effective for Open-Source Analytics with Apache Hadoop 3.0

- Apache Hadoop
- Apache Kafka



Azure Data Lake Storage [Gen2]

Designed for Big Data Analytics
File System Semantics & File Level Security for Scalability & Low-Cost

Spark Pools for Synapse

Parallel Processing Framework In-Memory to Boost Performance of Big-Data Analytic Applications

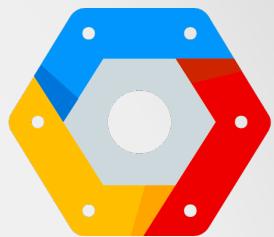


Azure Synapse Analytics

Fast, Flexible, & Secure Cloud Data Warehouse for Enterprises
SQL & PolyBase Features with Fast Loading Operations



Use-Case for Google Cloud Platform [GCP]



Google Pub/Sub

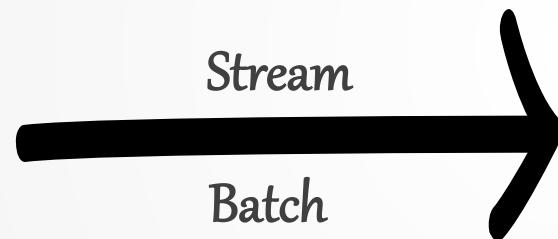
Global Messaging & Event Ingestion
Scale without Provisioning, Partitioning, or Load Isolation
Expand Pipelines to New Regions Simply with Global Topics

Cloud DataProc

Faster, Easier, Cost-Effective for Running Spark & Hadoop
Fast & Scalable Data Processing within 90 Seconds

Google BigQuery

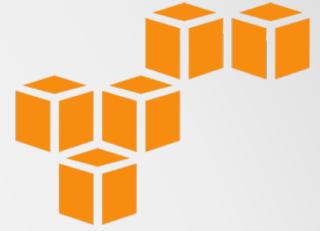
ServerLess [SaaS], Highly-Scalable, & Cost-Effective Cloud DW
In-Memory BI Engine & ML
Gartner 2019 – Magic Quadrant for Data Management Solutions



Google Cloud Storage [GCS]

Unified Object Storage for Developers & Enterprises
Optimize Price & Performance with 4 Storage Classes

Use-Case for Amazon AWS



Amazon Kinesis

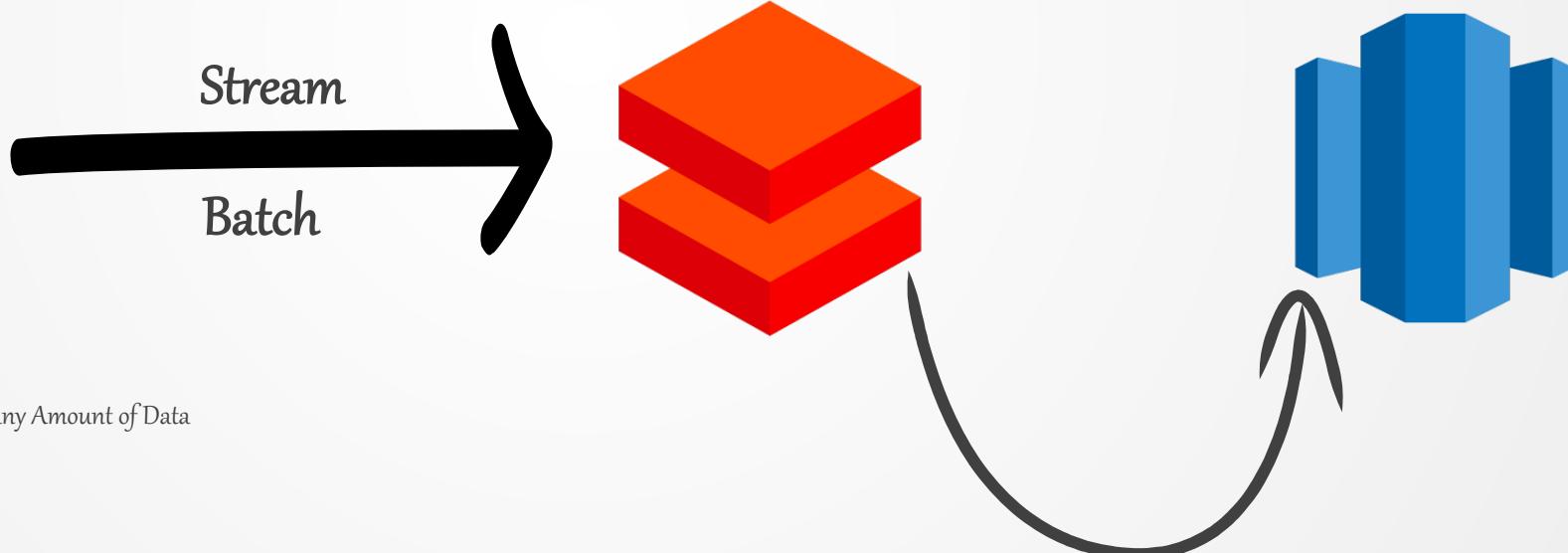
Easily Collect, Process & Analyze Streams in Real-Time

- Kinesis Data Streams
- Kinesis Data Firehose



Amazon S3

Object Storage Built to Store & Retrieve any Amount of Data
Storage Classes
Netflix & AirBnB

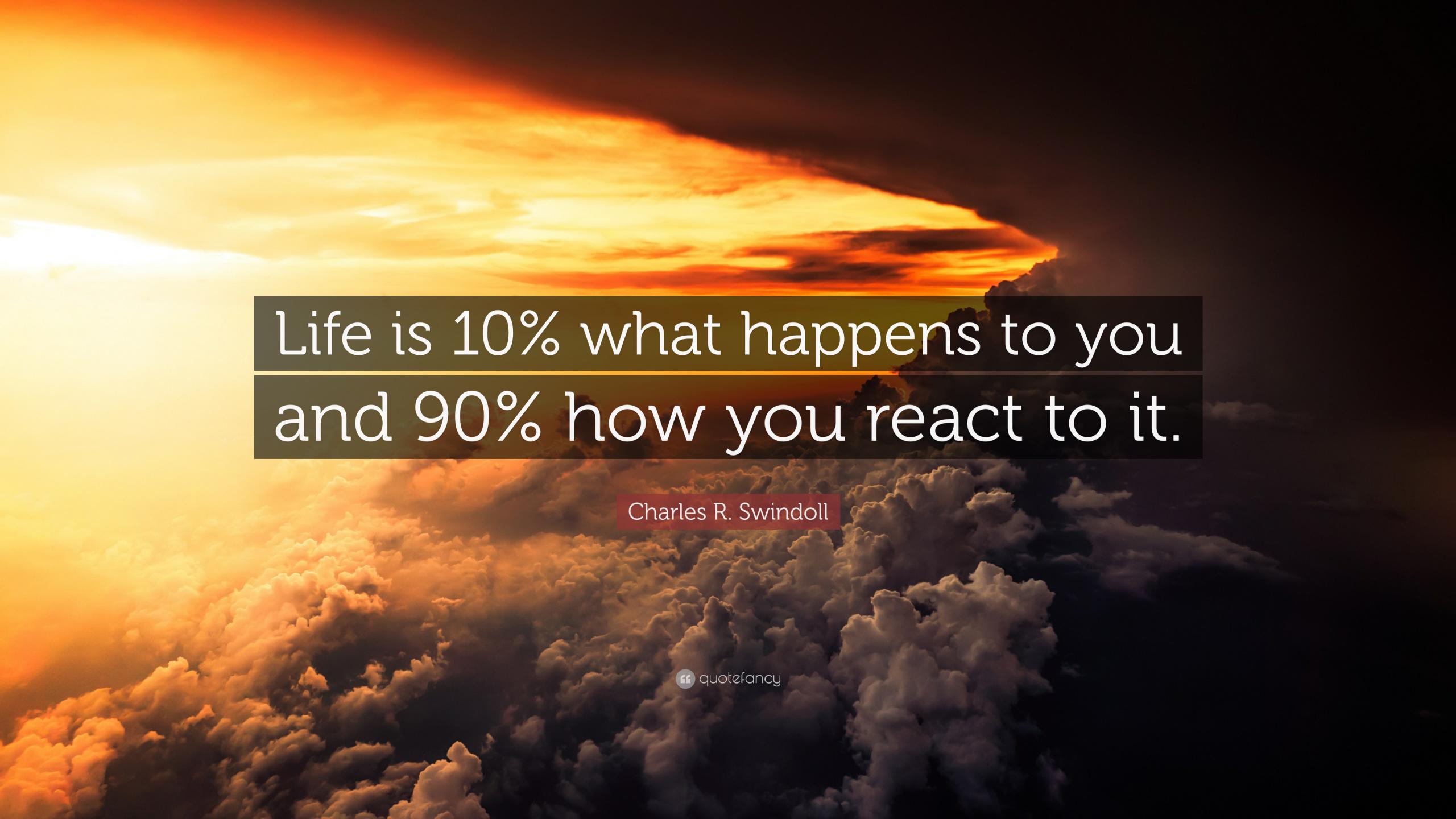


Databricks

Fast, Easy, & Collaborative Apache-Spark Analytics Service
Simplify Big Data & AI with Unified Analytics Platform

Amazon Redshift

Fast, Simple, Cost-Effective Modern Data Warehouse
MPP | ML | Result Caching & S3 Query Access



Life is 10% what happens to you
and 90% how you react to it.

Charles R. Swindoll

Apache Hive [The Godfather]



Open-Source Dw Offering

- Initially Developed by Facebook
- Used By Netflix



Analytics using SQL-Like

- SQL-Like Interface
- Processing Engine [MR | Tez | Spark]



Characteristics

- Analysis of Large Datasets
- Schema On-Read
- Structured & Unstructured Data
- Storages - HDFS | ADLS | WASB | S3 | GCS



Stinger Initiative

- 30/06/2014 by HortonWorks & Microsoft
- Human-Time [5-30 secs]



Optimizations



- ORC – Optimized Row Columnar
- LLAP [Live Long & Process]
- Sub-Second SQL Analytics with Intelligent Caching In-Memory



3.0

- Materialized Views
- Constraints & Default Values
- Apache Druid & Apache Kafka Connectors
- ACID v2 for Streaming Ingestion
- LLAP & Apache Spark Connector

Use-Case: Apache Hive

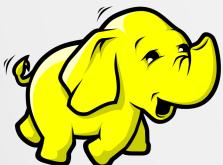
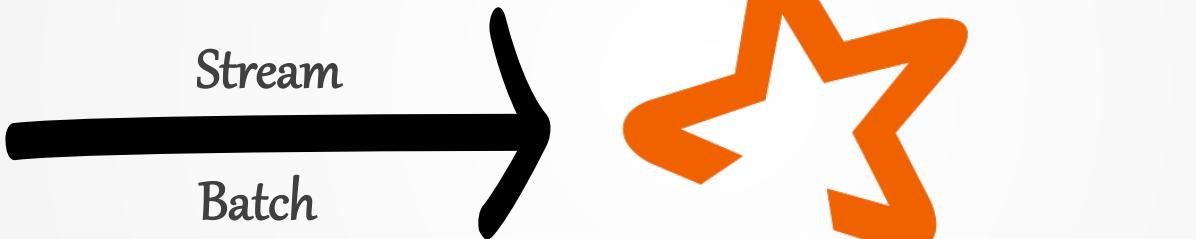


Apache Kafka

Distributed Streaming Platform
Real-Time Data Pipelines & Streaming Apps
Horizontally Scalable, Fault-Tolerant & Wicked Fast

Apache Spark

Unified Analytics Engine for Large-Scale Data Processing
Speed, Easy to Use, Generality & Runs Everywhere



HDFS

Hadoop Distributed File System
Run on Commodity Hardware
Designed for Large DataSets

Apache Hive

Data Warehouse [Dw] Open-Source with SQL-Like Interface
Hive LLAP – Sub-Second SQL Analytics with Intelligent Cache





Apache Hive [LLAP] as a Modern Data Warehouse [MDW]



Amazon Redshift



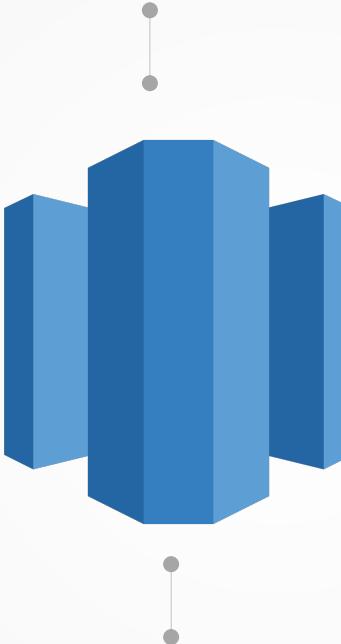
Cost-Effective

- Pay as you Go
- Predictable Cost
- Node Type



Amazon Redshift

- Most Popular Cloud Data Warehouse
- > 15K Customers using Amazon Redshift



- Petabyte-Scale
- Exabyte-Scale Data Lake Analytics with **Spectrum**
- Limitless Concurrency



Faster Performance

- Massively Parallel Processing
- Machine Learning
- Result Caching



Deploy & Manage

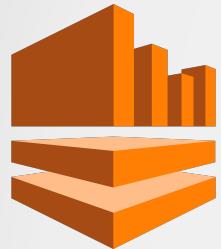
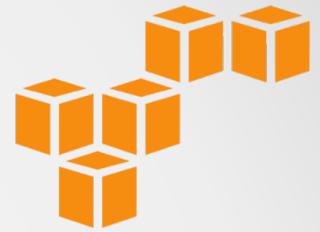
- Automated Provisioning
- Automated Backups
- Fault Tolerant
- Flexible Querying
- Third-Party Tools Integration



Integrations

- AWS S3 Data Lake
- AWS Glue
- Amazon Kinesis Firehose
- Amazon QuickSight
- Database Migration Service

Use-Case: Amazon AWS



Amazon Kinesis

Easily Collect, Process & Analyze Streams in Real-Time

- Kinesis Data Streams
- Kinesis Data Firehose

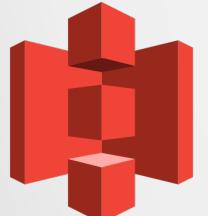
AWS Glue

Serverless Data Integration for Discover, Prepare and Combine Data for Analytics, ML & Application Development



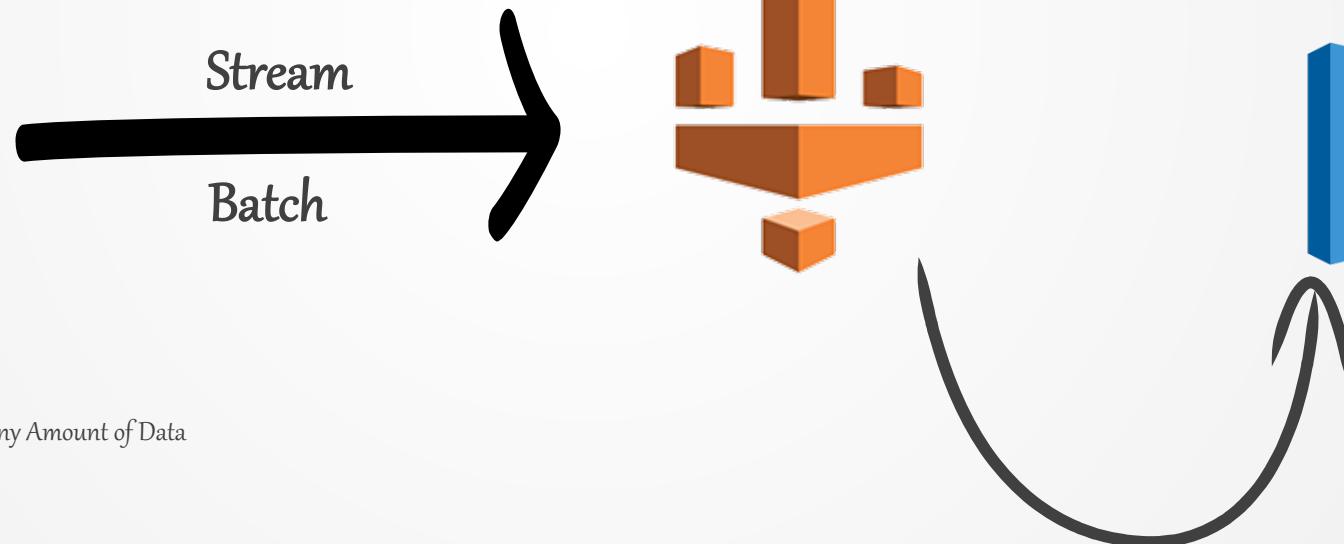
Amazon Redshift

Fast, Simple, Cost-Effective Modern Data Warehouse
MPP | ML | Result Caching & S3 Query Access



Amazon S3

Object Storage Built to Store & Retrieve any Amount of Data
Storage Classes
Netflix & AirBnB





Amazon Redshift as a Modern Data Warehouse [MDW]



Azure Synapse Analytics



Powerful Insights

Expand discovery of insights from all your data and apply machine learning models to all your intelligent apps



Azure SQL Data Warehouse [Evolved]

- Limitless Analytics Service
- Data Warehouse & Big Data Analytics



Limitless Scale

Deliver insights from all your data, across data warehouses and big data analytics systems, with blazing speed



Unified Experience

Significantly reduce project development time with a unified experience for developing end-to-end analytics solutions



Components

- SQL Analytics with SQL Pool & SQL on Demand
- Apache Spark Pools
- Data Integration & Studio



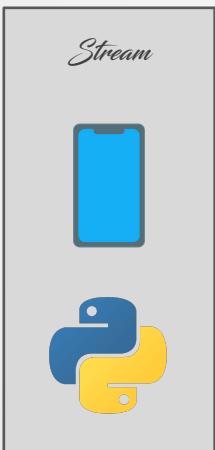
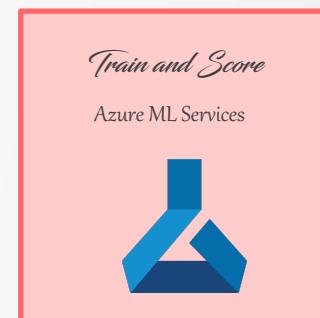
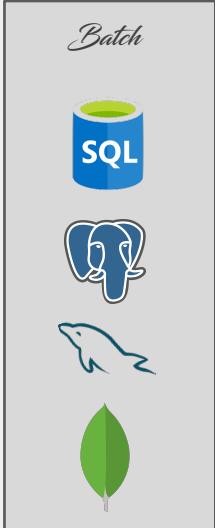
Features

- EDW - Azure SQL Dw Engine
- Data Lake Exploration
- Languages – T-SQL, Python, Scala, Spark SQL & .NET
- Orchestration – Azure Data Factory
- Streaming Ingestion & Analytics
- Integrated AI & BI

Big Data Architecture Proposal using Azure Synapse Analytics

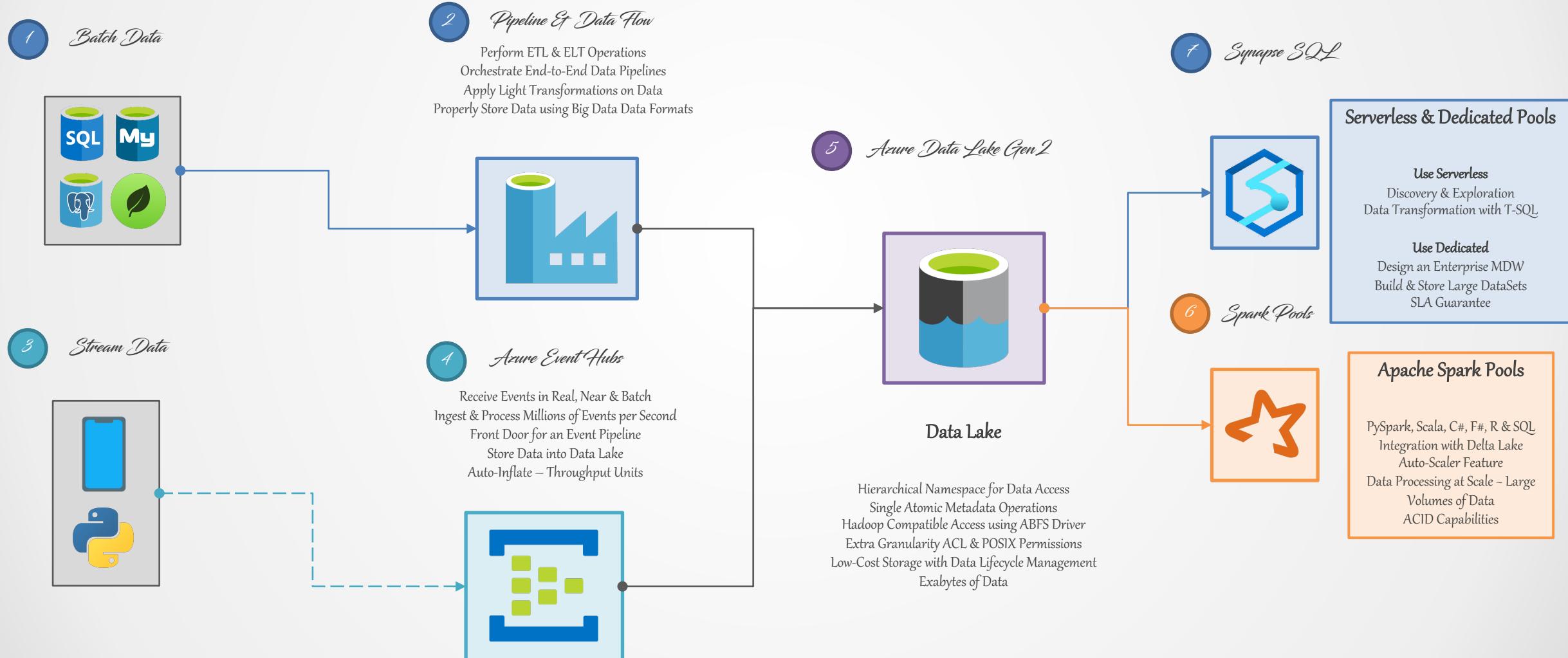


simplify the modern data warehouse creation by removing the glue required to integrate, secure and optimize the myriad of paas services that comprises the overall architecture



Use-Case: Microsoft Azure

Lambda Architecture





Azure Synapse Analytics as a Modern Data Warehouse [MDW]



Google BigQuery [The Kraken]

Leader for Data Management Solutions
for Analytics in 2019



Google's Serverless Offering

- Automatic Resource Provisioning
- SaaS Offering for Dw



Real-Time Analytics with SQL

- High-Speed Streaming Insertion API
- Standard ANSI:2011 SQL Support
- ODBC & JDBC Drivers



Big Data Ecosystem Integration

- Cloud DataProc
- Cloud DataFlow
- Apache Big Data Ecosystem
- Apache Hadoop & Apache Beam



Storage & Computation

- Separated Storage & Compute
- Choose Storage Tier
- Control Costs



Foundation for BI & AI



- EDW Google's Offering
- Integration, Transformation & Analyzes
- TensorFlow & BigQuery ML

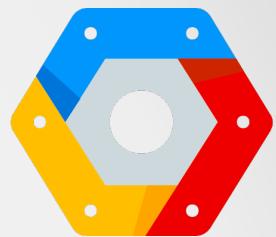


Programmatic Interaction

- REST API
- Java
- Python
- Node.js
- C#
- Ruby
- PHP

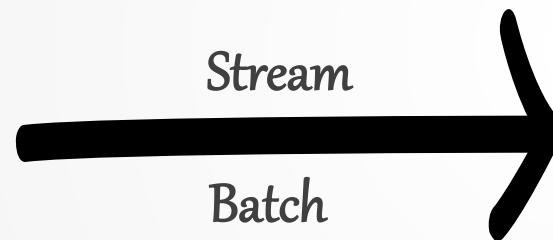


Use-Case: Google Cloud Platform [GCP]



Google Pub/Sub

Global Messaging & Event Ingestion
Scale without Provisioning, Partitioning, or Load Isolation
Expand Pipelines to New Regions Simply with Global Topics



Google Cloud Storage [GCS]

Unified Object Storage for Developers & Enterprises
Optimize Price & Performance with 4 Storage Classes

Cloud DataFlow

Simplified Stream & Batch Data Processing
Apache Beam [Java | Python | SQL]



Google BigQuery

ServerLess [SaaS], Highly-Scalable, & Cost-Effective Cloud DW
In-Memory BI Engine & ML
Gartner 2019 – Magic Quadrant for Data Management Solutions

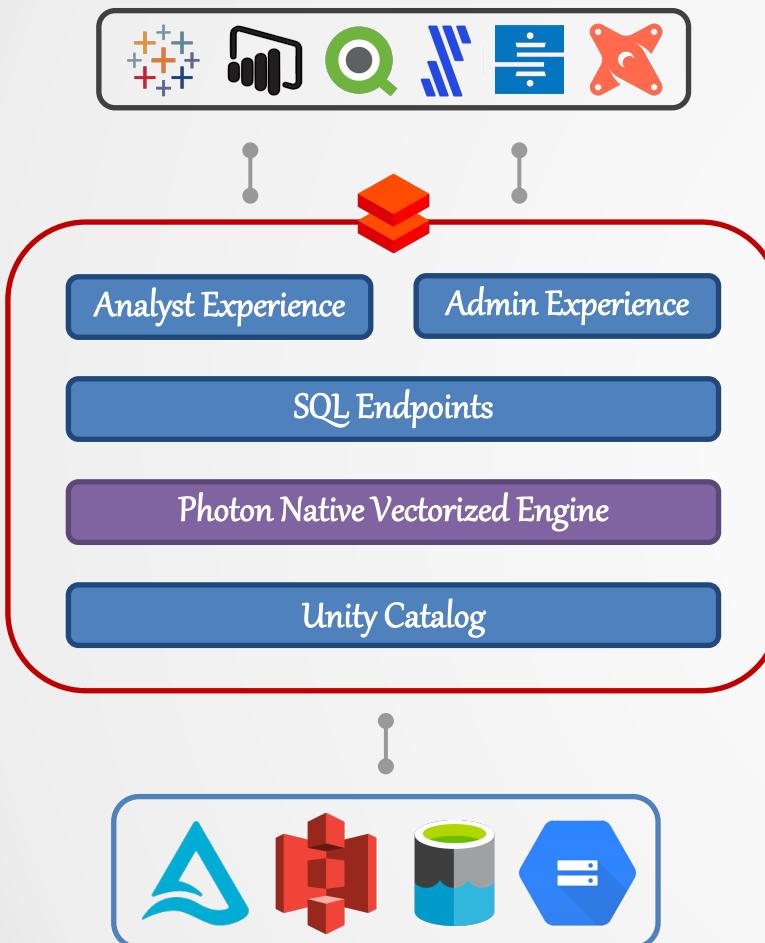




Big Query as a Modern Data Warehouse [MDW]



Databricks SQL



Integrations

Native Query and Visualization Tools, Provides Support of Existing BI Applications. Setting up Reliable Connections to Delta Lake Tables

Info

Real-World Performance & Easy to Use System for Analytics at Scale Based on SQL Workloads

Photon

Native Vectorized Engine Developed in C++ to Dramatically Improve Query Performance. Seamlessly Coordinate Work & Resources and Transparently Accelerate SQL Queries

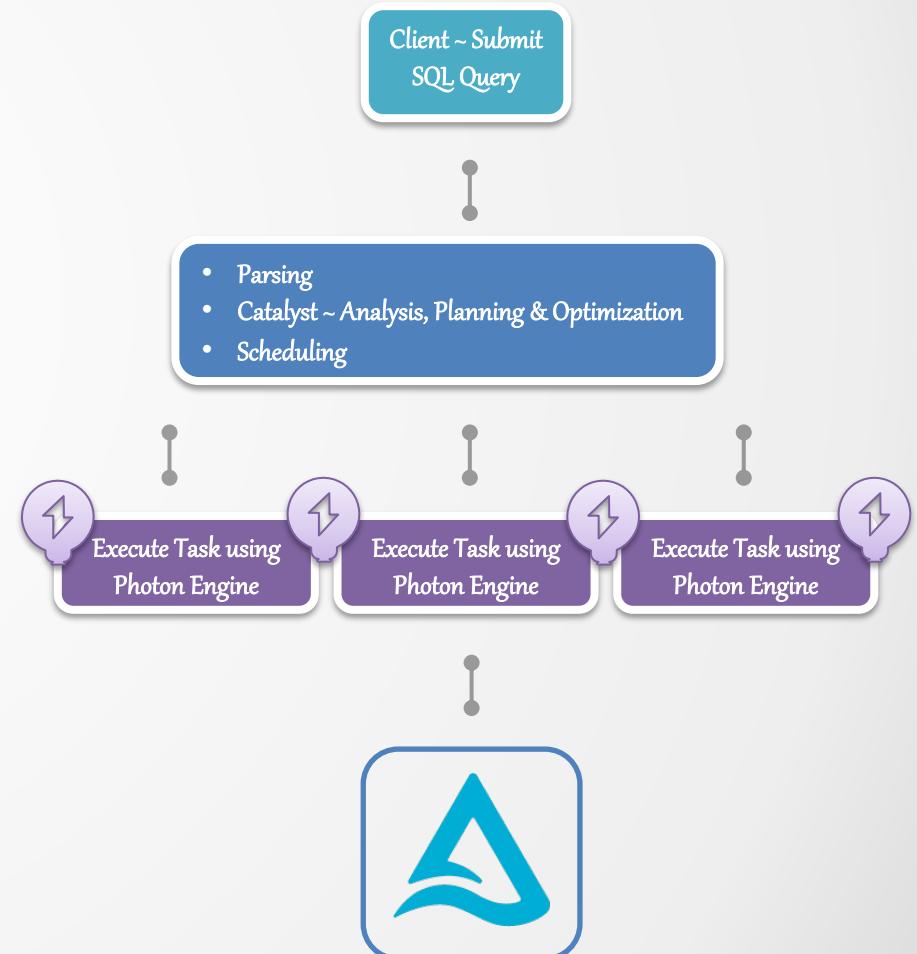
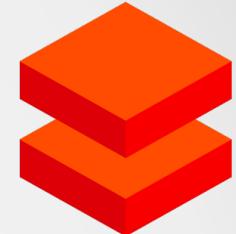
Unity Catalog

First Multi-Cloud Data Catalog Designed to Standardize Governance of Data & AI Assets on Data Lakehouse.

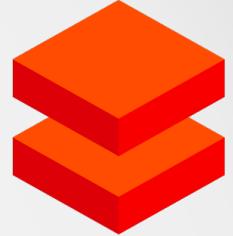
Curated Data

Structured, Semi-Structured, & Unstructured Data for Big Data Analytics. Open & Reliable Data Lake as Foundation

2x - 4x of Perf Improvement



Query Performance Comparison



30 TB TPC-DS Price and Performance Comparison

TPC-DS is an Enterprise-Class **Benchmark**, Published and Maintained By Transaction Processing Performance Council (TPC), Measure Performance of Decision Support Systems Running on SQL-Based Big Data Systems

Lower is Better



\$273



\$83



\$81



\$60



Databricks SQL as a Modern Data Warehouse [MDW]



The background of the image is a deep blue and purple starry night sky. At the bottom, there is a dark, silhouetted outline of what appears to be mountain peaks or hills.

**The only way of finding the limits
of the possible is by going
beyond them into the impossible.**

Arthur C. Clarke

Data Warehouse Dw - (1 Gen ~ 1992)



ETL for Data Centralization & BI Analysis

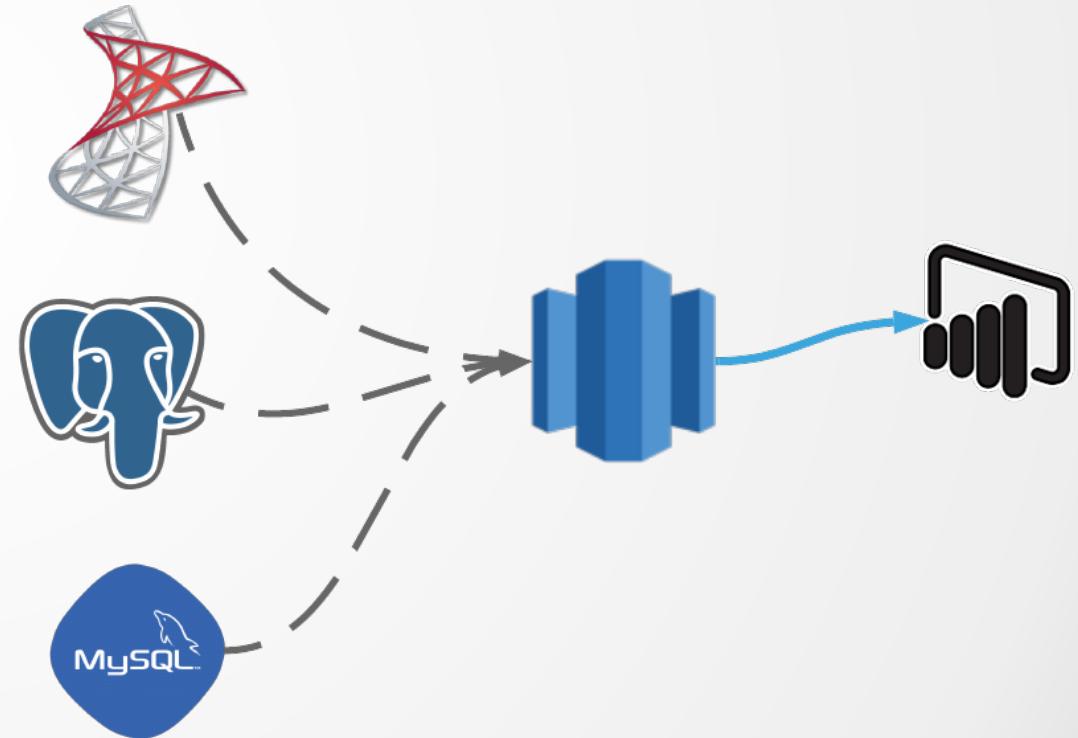
No Future Proof – Missing Predictions, Real-Time, Scale

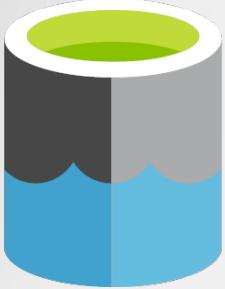


- Pristine
- Fast Queries
- Transactional



- Expensive for Scale, Not Elastic
- Require ETL, Stale Data, No Real-Time
- No Predictions, No ML
- Closed Formats [Lock In]





Hadoop Data Lake - (2 Gen ~ 2006)

ETL ALL Data, Scalable, Open Lake for ALL Use Cases

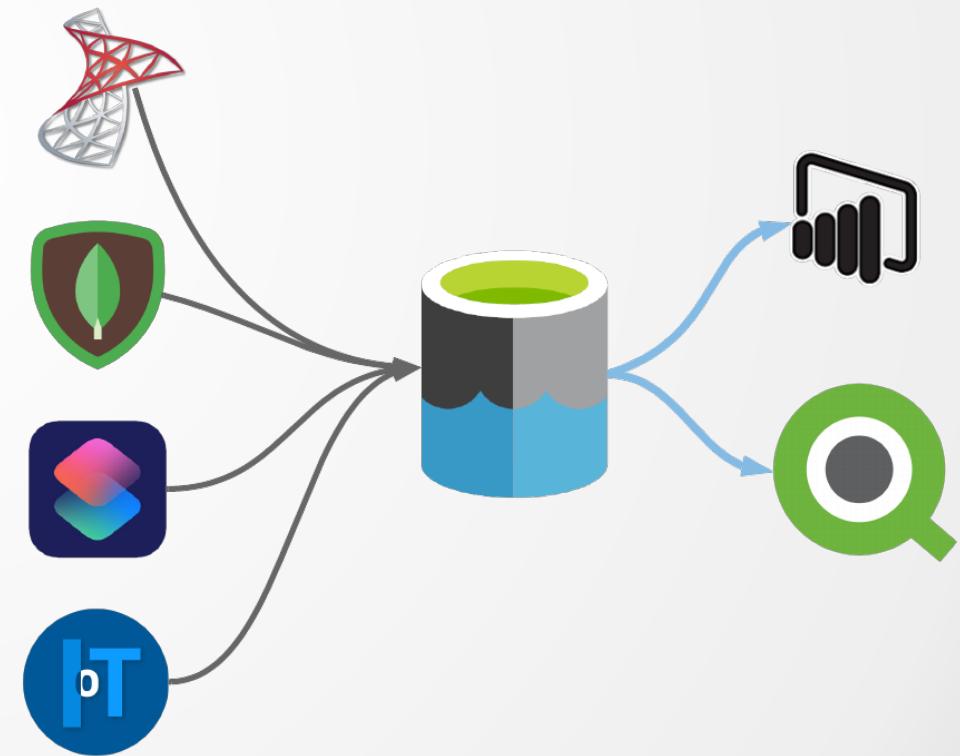
Become a Cheap Messy Data Store with Poor Performance



- Massive Scale
- Inexpensive Storage
- Open-Formats [Parquet, ORC]
- Promise of ML & Real-Time Streaming



- Inconsistent Data
- Unreliable for Analytics
- Lack of Schema
- Poor Performance





Data Lakehouse Delta Lake - (3 Gen ~ 2020)

A Unified Data Management System for Real-Time Big Data
Powerful Transactional Storage Layer

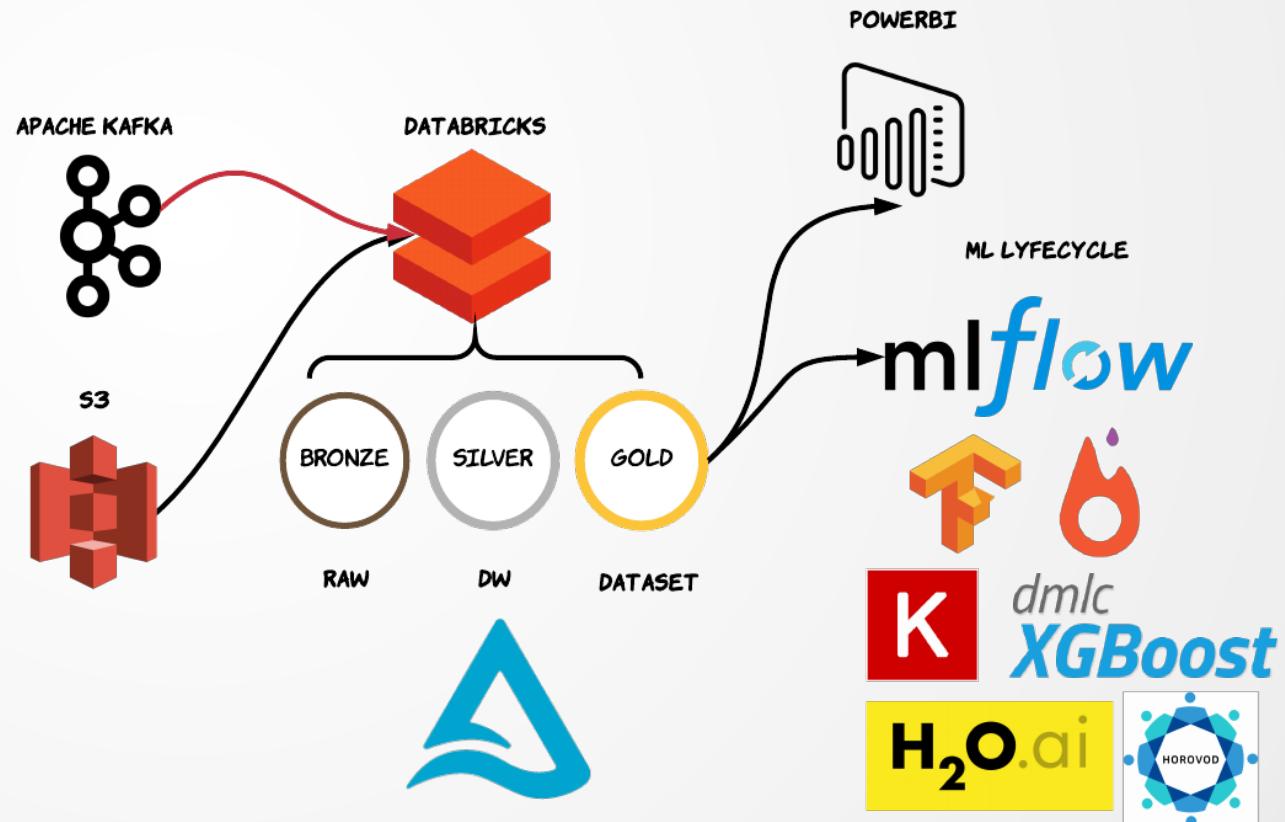


- The Good of Dw
- The Good of Data Lakes
- Decoupled Compute & Storage
- ACID Transactions & Data Validation
- Data Indexing & Caching [10x ~100x]
- Real-Time Streaming Ingest

Key Features



- ACID
- Scalable Metadata
- Time Travel
- Open Format
- Batch & Streaming Source & Sink
- Schema Enforcement & Evolution
- Audit History
- Insert, Updates, Delete & Merge



Use-Case: Open-Source Software [OSS] with Delta Lake



Apache Kafka

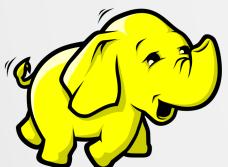
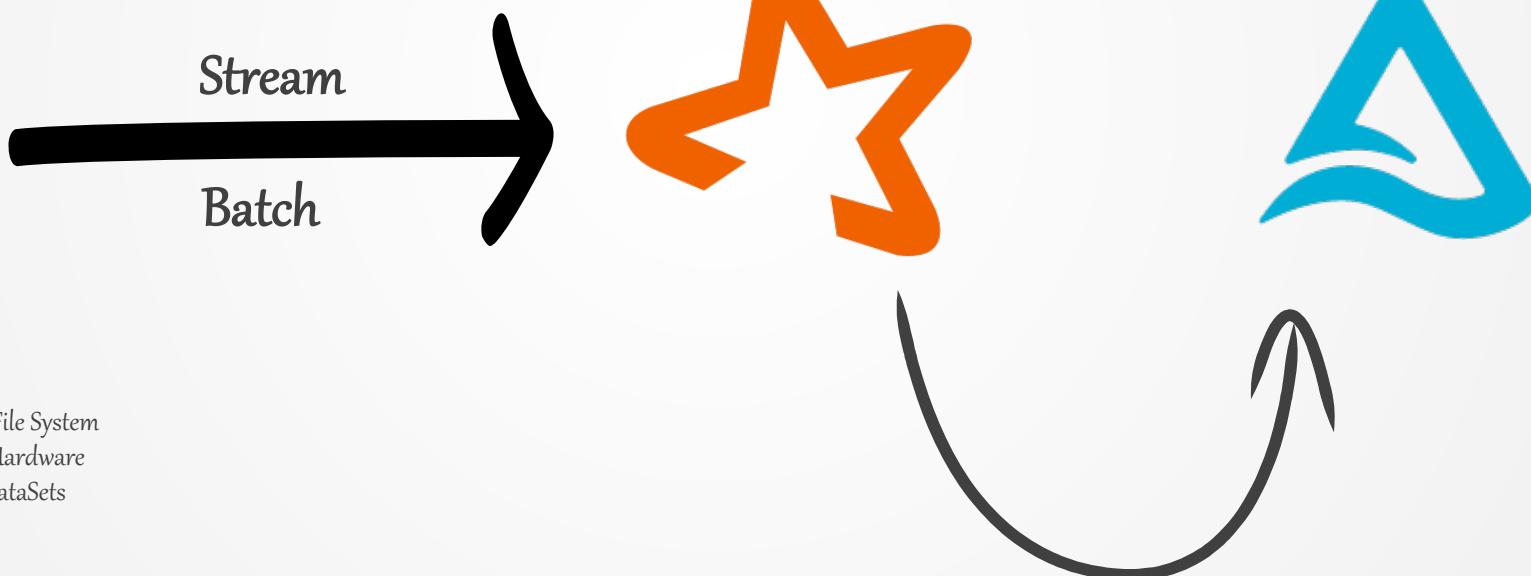
Distributed Streaming Platform
Real-Time Data Pipelines & Streaming Apps
Horizontally Scalable, Fault-Tolerant & Wicked Fast

Apache Spark

Unified Analytics Engine for Large-Scale Data Processing
Speed, Easy to Use, Generality & Runs Everywhere

Delta Lake

Open-Source Storage Layer with ACID Capabilities
Batch & Streaming Unified



HDFS

Hadoop Distributed File System
Run on Commodity Hardware
Designed for Large DataSets



Data Lakehouse using Delta Engine for Data Analytics at Scale



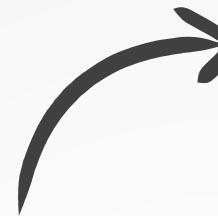
Ignore the noise,
focus on your work.

Anonymous



Data Virtualization

Access Data in Any Source, Shape or Structure without Copying from Source



Processing Query Engine

Connects, Discover, Plan and Analyze Different Data Stores for Query at Scale



Modern Data Warehouses

Redshift, BigQuery



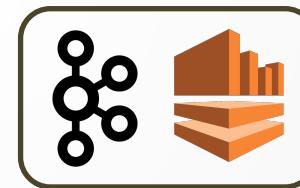
Real-Time OLAP

Apache Druid & Apache Pinot



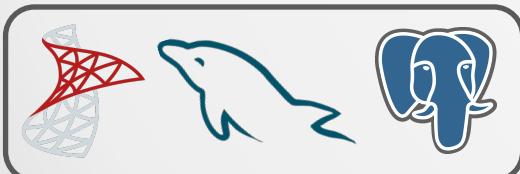
Real-Time Ingestion

Apache Kafka, Kinesis



Relational Databases

SQL Server, MySQL, Postgres



NoSQL

Cassandra, MongoDB, Redis, ElasticSearch



Data Lake

HDFS, S3, Blob Storage, GCS & MinIO [S3*]



Virtualization Engines

Virtualization Engines for Querying Data at Scale



PolyBase

Transact-SQL Queries ~ Reads Data from External Sources

- SQL Server
- Oracle
- Teradata
- MongoDB
- HDFS
- Blob Storage

Microsoft Azure



Amazon Athena

Serverless, Interactive Query Service to Query Data and Analyze Big Data in Amazon S3 using Standard SQL

AWS



Redshift Spectrum

Query Data Directly from Amazon Redshift from Files on Amazon S3

AWS



Trino



Query Engine Runs at Ludicrous Speed.
Fast Distributed SQL Query Engine for Big Data Analytics

Open Source



Dremio



Next-Generation Data Lake Engine for Interactive Queries Directly from Cloud Data Storage

- ADLS
- S3
- HDFS
- MongoDB
- SQL Server
- ElasticSearch

Open-Source

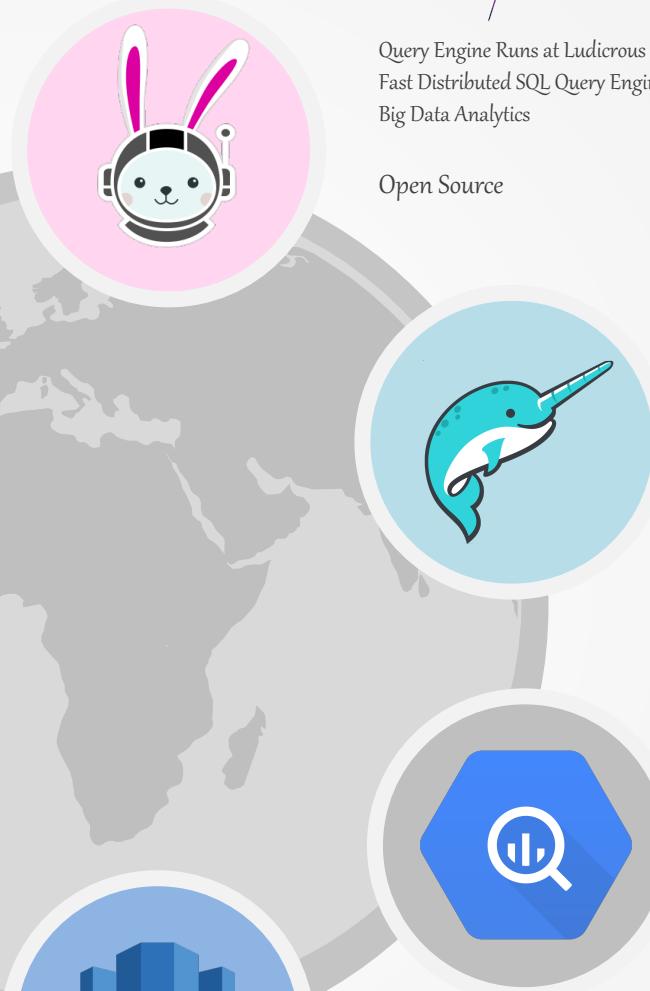


Google BigQuery

Query External Data Sources [Federated] Directly using a Table Reference

- Cloud BigTable
- Cloud Storage
- Google Drive
- Cloud SQL
- Omni

Google Cloud Platform [GCP]



Trino at Ludicrous Speed

Fastest Bunny on Forest, Eating Any Source of Data



Trino

- Speed & Scale
- Simplicity
- Versatile
- In-Place Analysis
- Query Federation
- Runs Everywhere
- Trusted
- Trino Software Foundation [OSS]



Designed For

- Access Data using SQL
- Data Warehousing
- Analytics
- OLAP



Trino is a Tool Designed to Efficiently Query
Vast Amounts of Data using Distributed
Queries in a Terabytes & Petabytes Range



Features

- Memory Management & Spilling using Compression and Encryption
- Resource Groups
- Distributed Sort
- Dynamic Filtering ~ Query Performance, Network Traffic Reduction & Remote Load Reduction
- Cost Based Optimization - CBO
- Query Pushdown



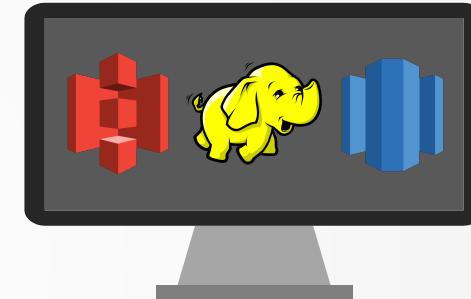
Apache ORC ~ Optimized Row Columnar

- Smallest, Fastest Columnar Storage for Apache Hadoop Workloads
- ACID Support
- Built-In Indexes
- Complex Types
- Integration with Apache Spark & Trino



Query Execution Model

Turn SQL Statements into Queries,
Executes Across a Distributed
Cluster of Coordinator and Workers



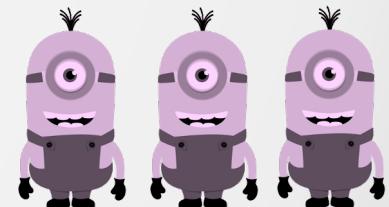
Coordinator a.k.a The Brain

Parsing Statements, Planning Queries, &
Managing Trino Worker Nodes



Workers a.k.a The Slave

Responsible for Executing Tasks & Processing
Data. Fetch Data from Connectors and
Exchange Intermediate Data in Parallel



Distributed SQL Query Engine for DW & Big Data Analytics ~ Data Virtualization





ONE WAY
SOLUTION