ONEWAY

SOLUTION

One Way Solution

Data Warehouse [Dw]

Data Engineering – [Day 4]

LUAN MORENO

CEO & CDO

Data Engineer & Data Platform MVP

Confluent Certified Developer for Apache Kafka [CCDAK]

Near Real-Time ETL

Test Data Engineering – [Day 3]

Data Engineering — Near Real-Time ETL

3

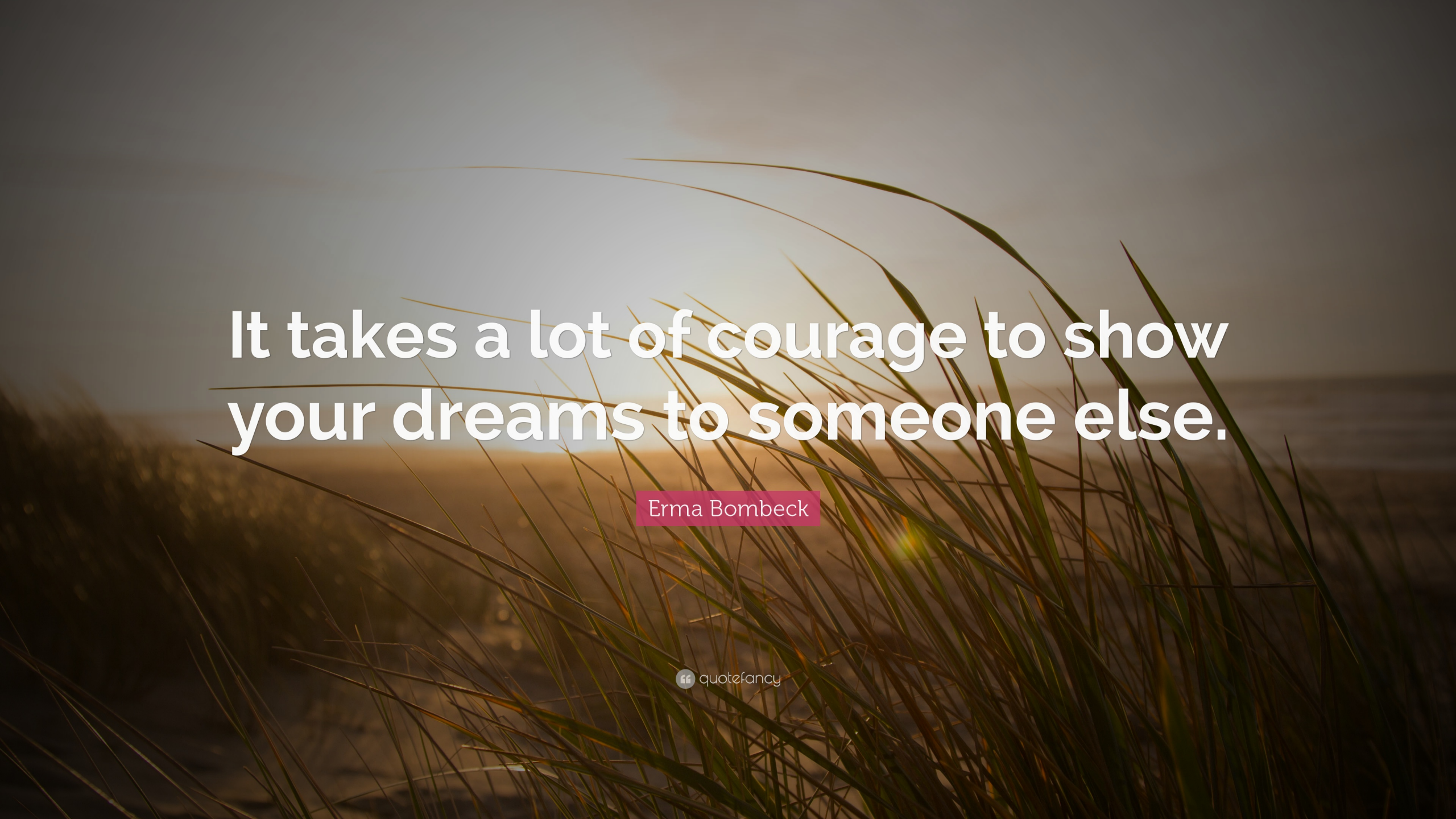
Event Stream Processing
Real-Time Stream Processing Engines
Python & Streaming Engines
StreamingSQL
Apache Kafka
Spark Streaming & Structured Streaming
Use-Cases

Agenda

4

ETL vs. ELT
TDW vs. MDW vs. Lakehouse
Use-Cases – AWS, Azure & GCP
Azure SQL Database
Azure Synapse Analytics
Amazon Redshift
Google BigQuery
Apache Hive
Apache Druid
Delta Lake

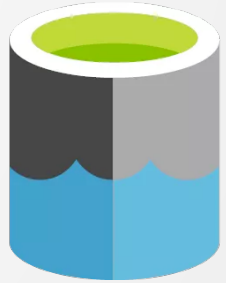
One Way Solution


It takes a lot of courage to show your dreams to someone else.
Erma Bombeck
quotefancy

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**.

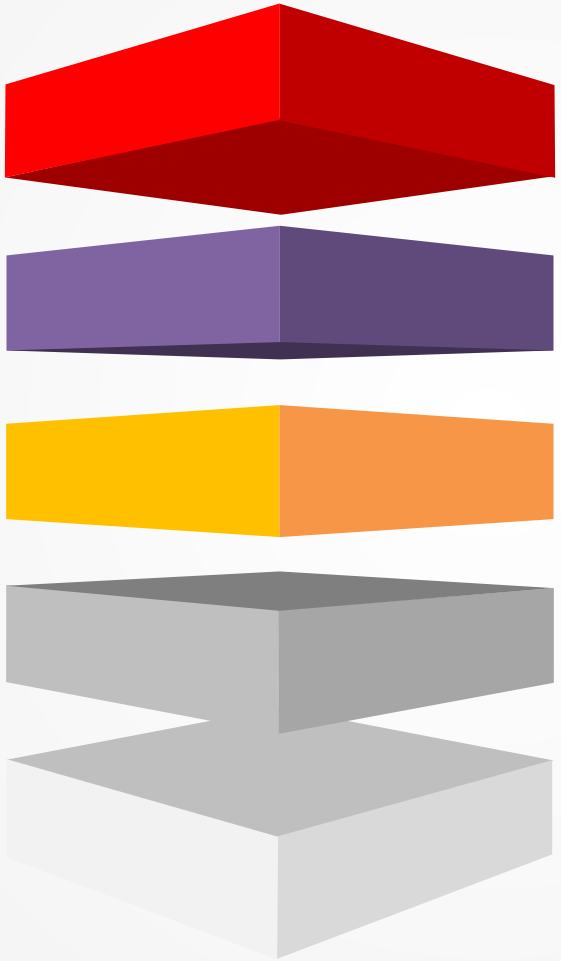
Extract/Load/Transform (ELT) similarly extracts data from one or multiple remote sources, but then loads it into the target **Data Lake** without any other formatting. The transformation of data, in an ELT process, happens within the target database. ELT asks less of remote sources, requiring only their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



Traditional Data Warehouse [Dw]

History

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



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

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]

Star Schema Model

Model

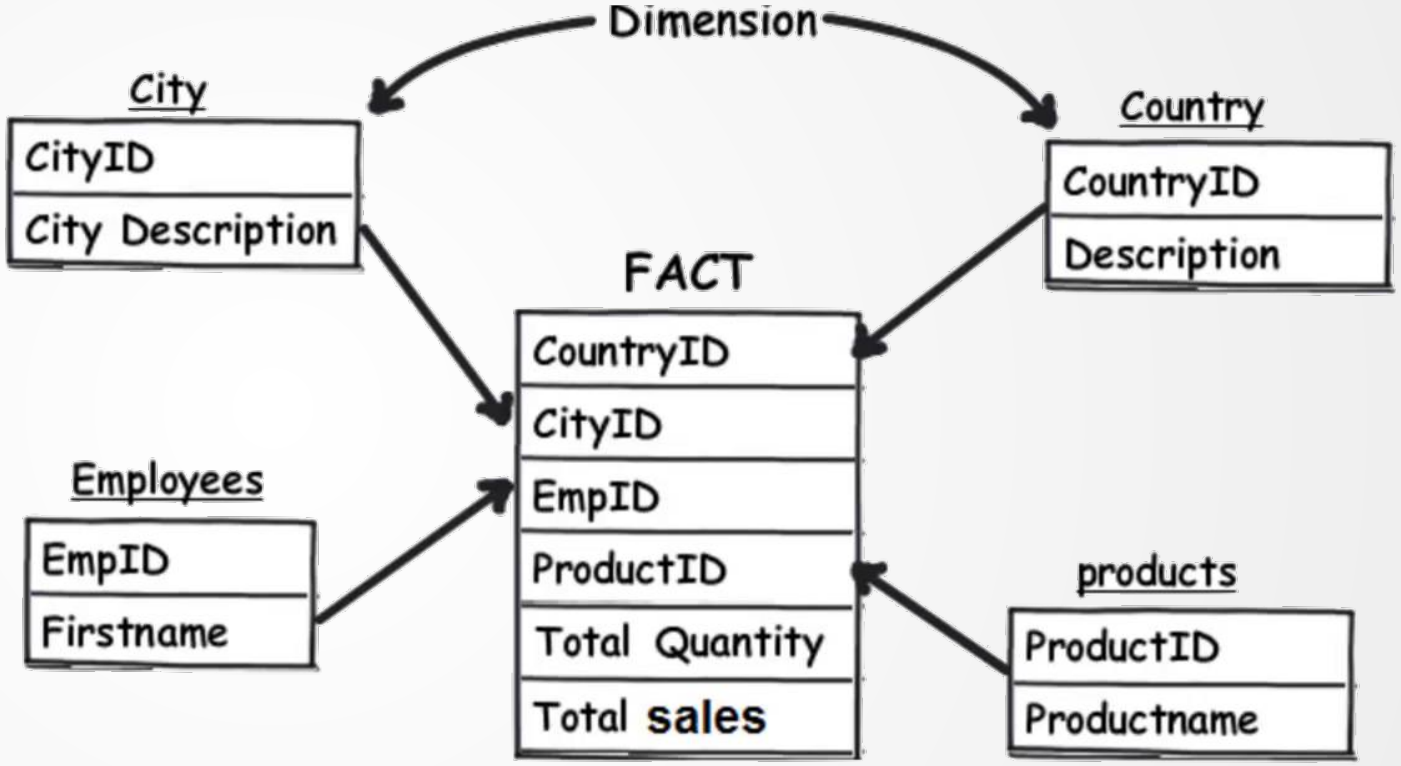
- Separates Business Process Data into **Facts**
- **Dimensions** with Descriptive Attributes
- Measurable & Quantitative Data

Benefits

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

Disadvantages

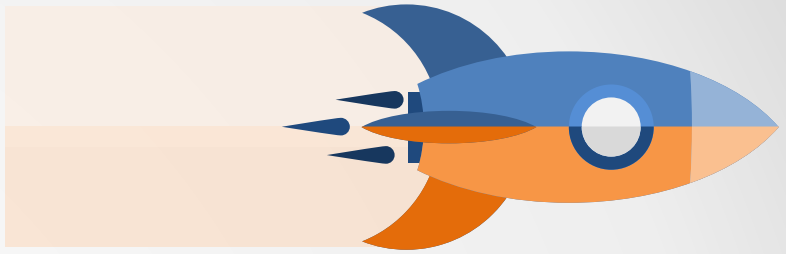
- Data Integrity [De-Normalized State]
- Batch Processing Load Fashion



Doubt kills more dreams than failure ever will.

Suzy Kassem

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



Modern Data Warehouse – [DW]

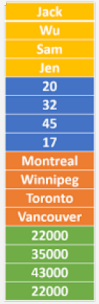
- Analytics Platform for Enterprises
- Scalability – Horizontally vs. Vertically
- PaaS – $ per Hs & SaaS – $ per Query
- SQL-Like Interface [SQL]

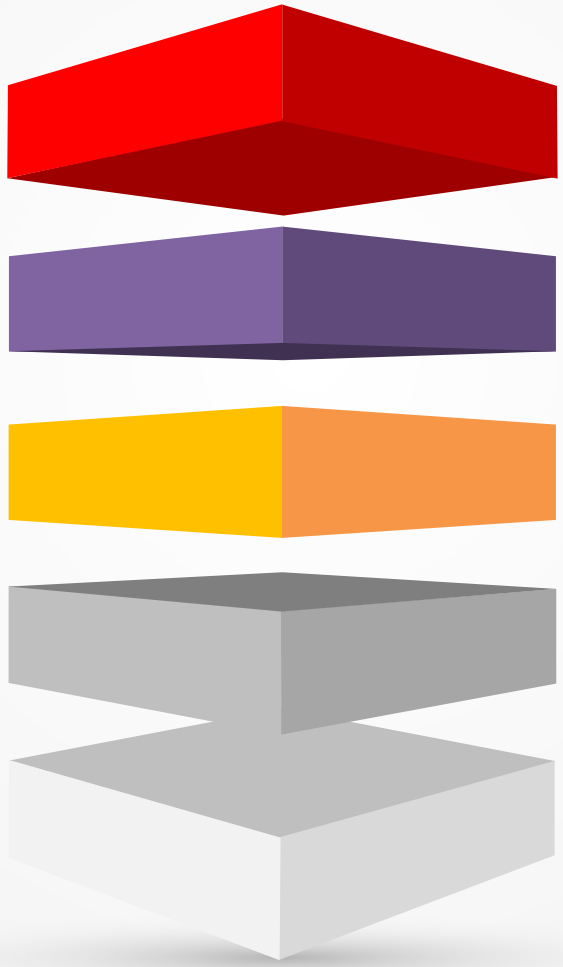
Physical Hardware

- Massively Parallel Processing [MPP]
- "Loosely Coupled" or "Shared Nothing"

Columnar Data Storage Type

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





Cloud-Based Data Warehouses Systems

- Amazon Redshift
- Azure SQL Data Warehouse
- Google BigQuery
- Snowflake

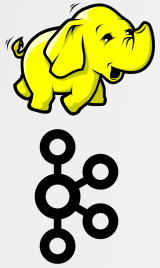
Elastic Compute & Storage

- [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

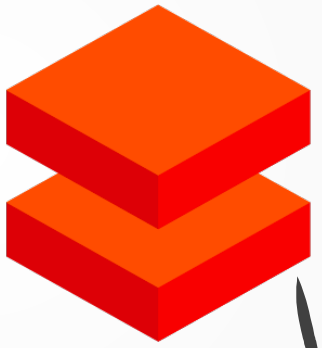
- Apache Hadoop
- Apache Kafka

Stream

Batch

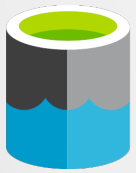
Azure Databricks

Fast, Easy, & Collaborative Apache-Spark Analytics Service
Azure Databricks & Azure SQL Dw with PolyBase



Azure Synapse Analytics

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



Azure Data Lake Storage [Gen2]

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

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

Google Cloud Storage [GCS]

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

Stream

Batch

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

Use-Case for Amazon AWS

Amazon Kinesis

Easily Collect, Process & Analyze Streams in Real-Time

- Kinesis Data Streams
- Kinesis Data Firehose

Databricks

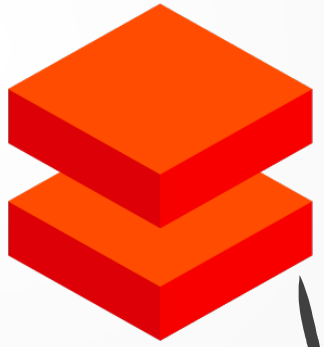
Fast, Easy, & Collaborative Apache-Spark Analytics Service
Simplify Big Data & AI with Unified Analytics Platform

Amazon Redshift

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

Stream

Batch



Amazon S3

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

Use-Case for [OSS] – Open-Source Platform

Apache Kafka

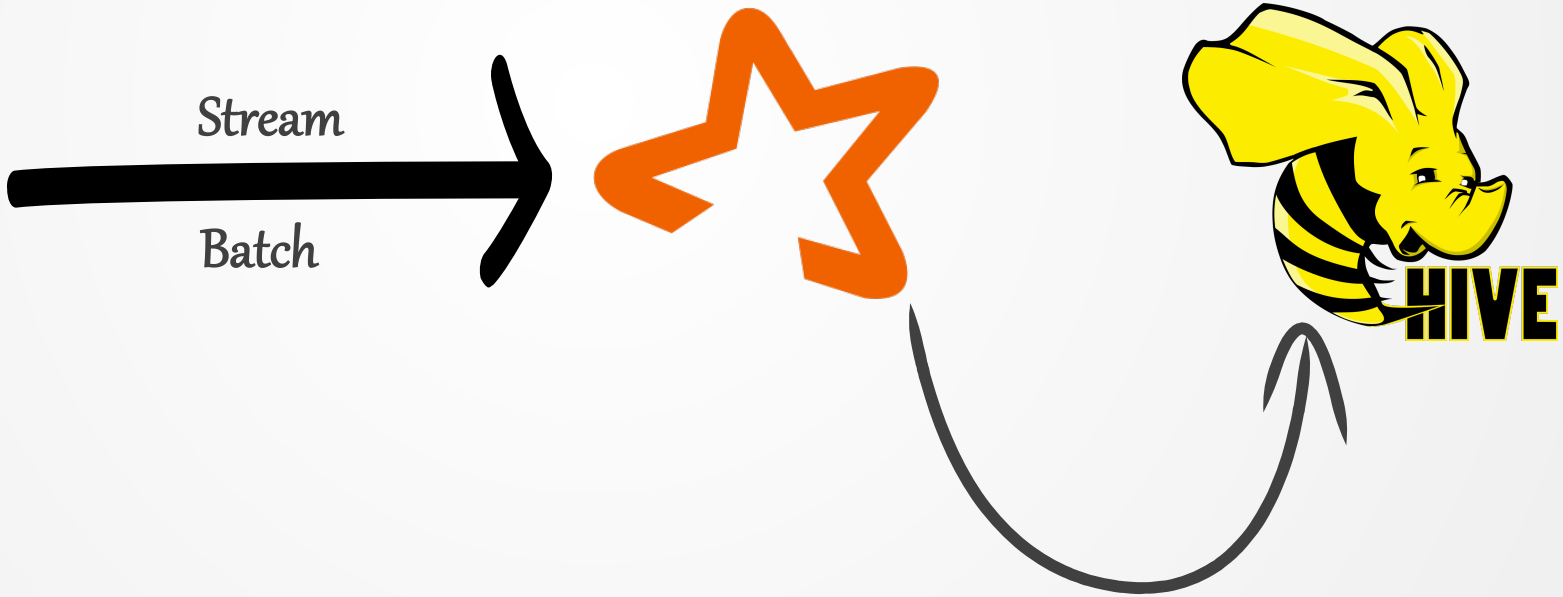
Distributed Streaming Platform
Real-Time Data Pipelines & Streaming Apps
Horizontally Scalable, Fault-Tolerant & Wicked Fast

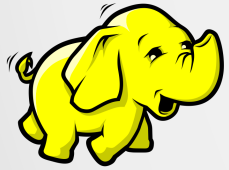
Apache Spark

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

Apache Hive

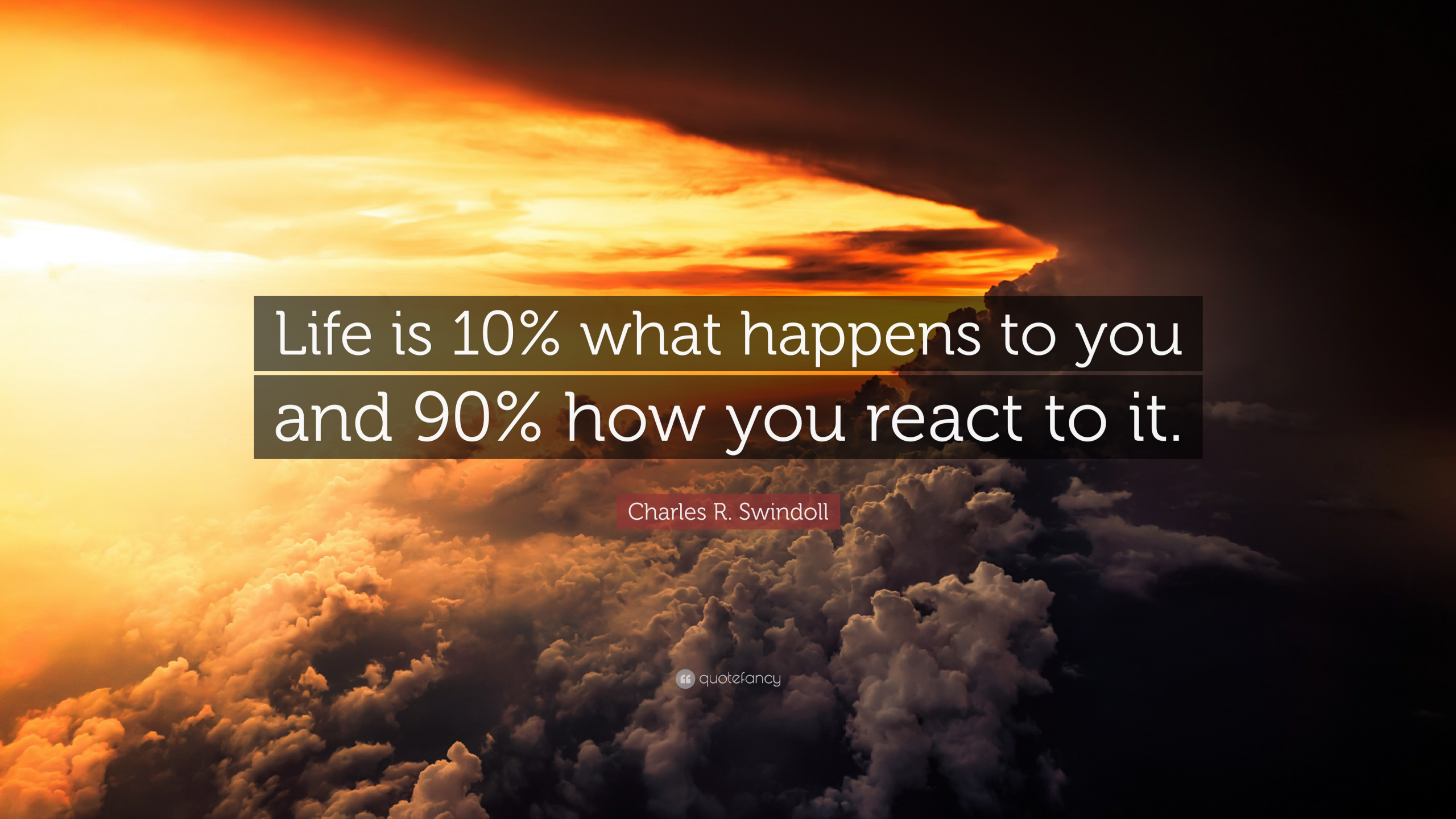
Data Warehouse [Dw] Open-Source with SQL-Like Interface
Hive LLAP – Sub-Second SQL Analytics with Intelligent Cache





HDFS

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


Life is 10% what happens to you
and 90% how you react to it.
Charles R. Swindoll
quotefancy

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 Data Warehouse


The past has no power over the present moment.
Eckhart Tolle
quotefancy

Azure Synapse Analytics

Formerly Azure SQL DW

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

Components

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

Limitless Scale

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

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

Unified Experience

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

Use-Case: Microsoft Azure

Azure Event Hubs

Simple, Secure, & Scalable Real-Time Data Ingestion

- AMQP
- HTTPS
- Apache Kafka

Azure Data Lake Storage [Gen2]

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

Stream

Batch

Azure Stream Analytics

Serverless Real-Time Analytics [SQL Interface]
Machine Learning

Azure Synapse Analytics

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

Azure Synapse Analytics as a Data Warehouse


Either you run the day
or the day runs you.
Jim Rohn
quotefancy

Amazon Redshift

Cost-Effective

- Pay as you Go
- Predictable Cost
- Node Type

Amazon Redshift

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

Scalability

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

Faster Performance

- Massively Parallel Processing
- Machine Learning
- Result Caching

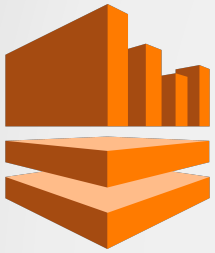
Integrations

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

Deploy & Manage

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

Use-Case: Amazon AWS



Amazon Kinesis

Easily Collect, Process & Analyze Streams in Real-Time

- Kinesis Data Streams
- Kinesis Data Firehose

Amazon Kinesis Data Analytics

Analyze Streaming Data
Query Streams of Data

Amazon Redshift

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

Stream

Batch

Amazon S3

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


The best revenge is
massive success.
Frank Sinatra
quotefancy

Google BigQuery [The Kraken]

Leader for Data Management Solutions
for Analytics in **2019**

Storage & Computation

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

Google's Serverless Offering

- Automatic Resource Provisioning
- SaaS Offering for Dw

Foundation for BI & AI

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

Real-Time Analytics with SQL

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

Programmatic Interaction

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

Big Data Ecosystem Integration

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

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

Stream

Batch

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


Always turn a negative situation
into a positive situation.
Michael Jordan
quotefancy

Apache Hive [The Godfather]

Stinger Initiative

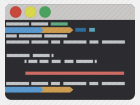
- 30/06/2014 by HortonWorks & Microsoft
- Human-Time [5-30 secs]

Open-Source Dw Offering

- Initially Developed by Facebook
- Used By Netflix

Optimizations

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



Analytics using SQL-Like

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

3.0

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

Characteristics

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

When you want something, all the universe conspires in helping you to achieve it.

Paulo Coelho

Apache Druid



Deploy

- AWS | GCP | Azure
- Kubernetes

Open-Source Data Store Offering

- Sub-Second Queries
- Historical & Real-Time data

Apache Hive

- Indexing Complex Query Results
- SQL Interface for Apache Druid
- Execute Complex Queries
- Efficient Execution Layer

Modern Cloud-Native Analytics DB

- Next-Gen Open Source Alternative for Analytical Database
- Vertica | Greenplum | Exadata
- Snowflake | BigQuery | Redshift

Use Cases

- Streaming & Operational Data [Apache Kafka]
- OLAP & BI
- User Activity & Behavior
- Network Flows
- Digital Marketing
- IoT & Device Metrics

Easy Integration

- Apache Kafka
- Amazon Kinesis
- HDFS
- Amazon S3

Use-Case: Open-Source Software [OSS]

Apache Kafka

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

Apache Spark

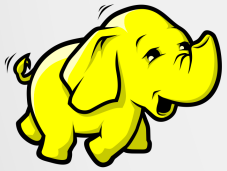
Unified Analytics Engine for Large-Scale Data Processing
Speed, Easy to Use, Generality & Runs Everywhere

Apache Hive

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

Stream

Batch



HDFS

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

Apache Druid

High Performance Real-Time Analytics Database
Column-Oriented Storage with SQL Query [Apache Calcite]


The only way of finding the limits of the possible is by going beyond them into the impossible.
Arthur C. Clarke
quotefancy

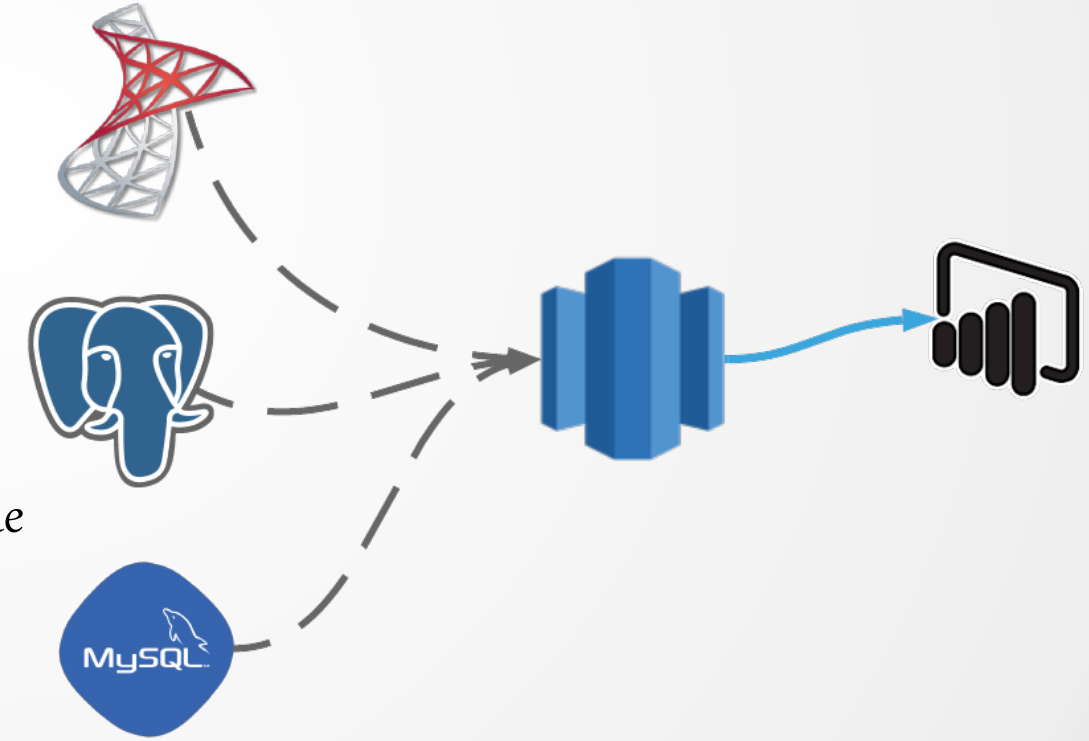
Data Warehouse [Dw] – (1 Gen ~ 1992)

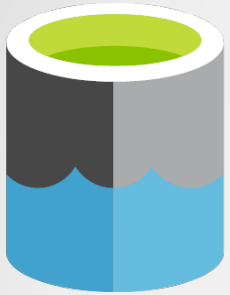
ETL for Data Centralization & BI Analysis

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]





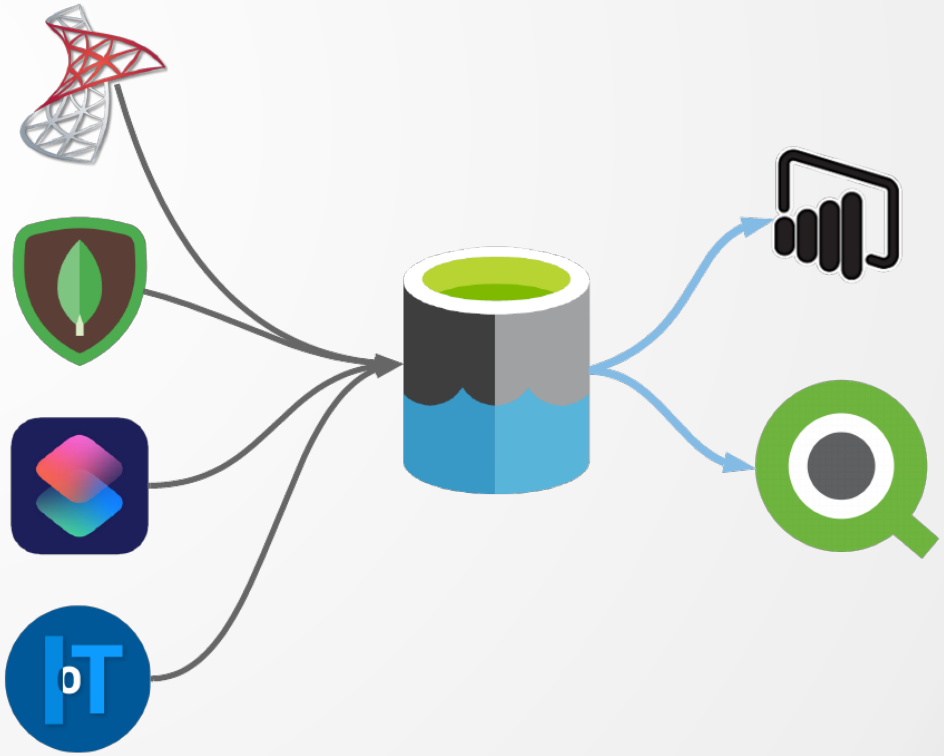
Hadoop Data Lake – (2 Gen ~ 2006)

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

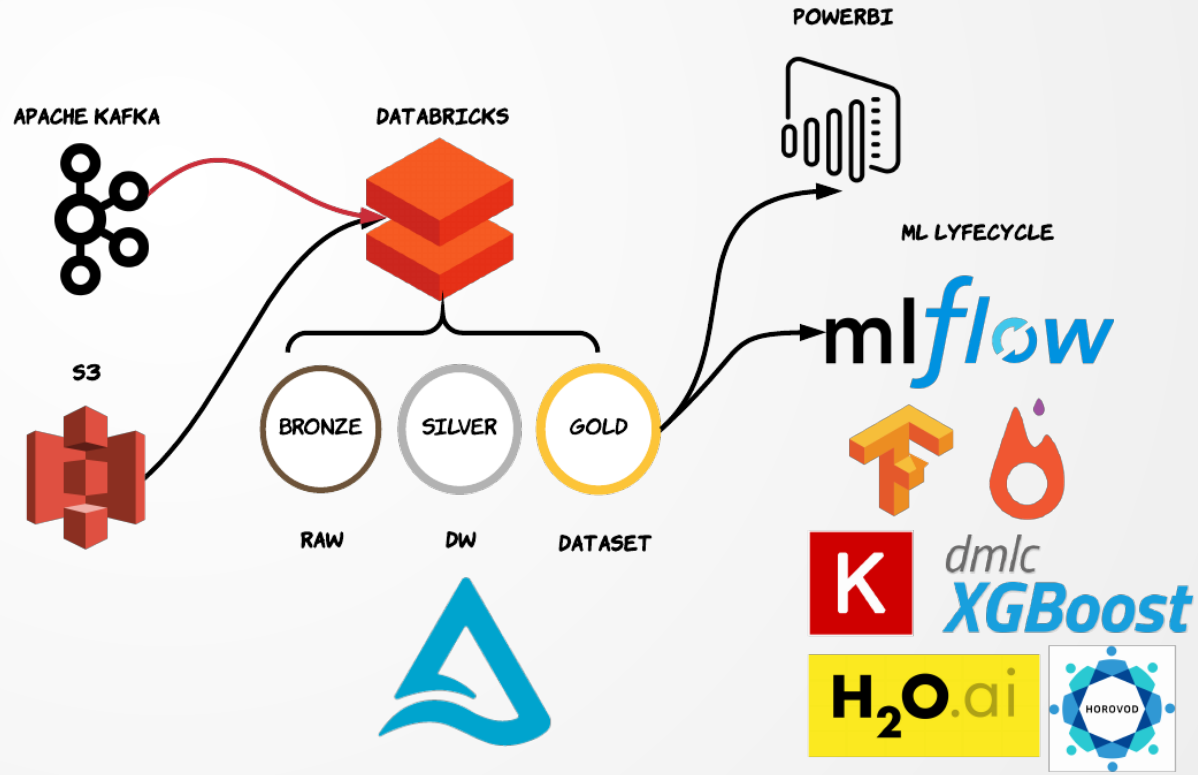


Data Lakehouse [Delta Lake] – (3 Gen ~ 2020)

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
- Updates & Deletes

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

Apache Kafka

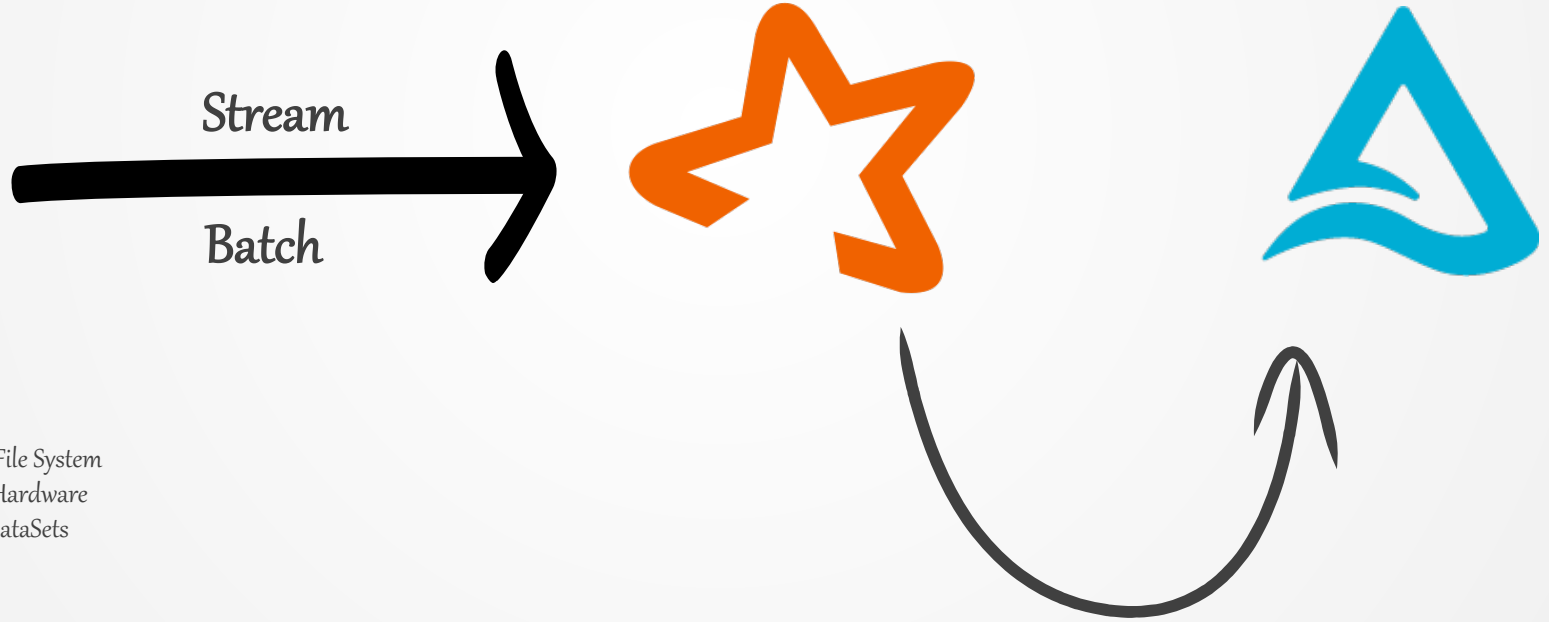
Distributed Streaming Platform
Real-Time Data Pipelines & Streaming Apps
Horizontally Scalable, Fault-Tolerant & Wicked Fast

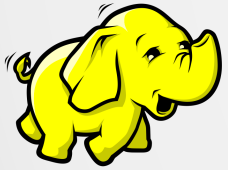
Apache Spark

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

Delta Lake as a Data Warehouse

ONEWAY

SOLUTION