ONEWAY

SOLUTION

One Way Solution

Foundation, Apache Spark & Databricks

Data Engineering – [Day 1]

LUAN MORENO

CEO & CDO

Data Engineer & Data Platform MVP

Confluent Certified Developer for Apache Kafka [CCDAK]

Agenda

1

Big Data
Cloud Computing
Apache Hadoop
Big Data as a Service [BDaaS]
Apache Fundamentals
Apache Spark 3.0 Features
Databricks
Big Data Architectures

One Way Solution


Do not fear mistakes. You will know failure. Continue to reach out.
Benjamin Franklin
quotefancy

Big Data

Netflix, Inc.

137 million users worldwide with consumption of 25% of the world's internet bandwidth

Spotify

191 million users worldwide with more than 30 million of songs available

Airbnb, Inc.

income of $107 millions with average of 2m+ of people staying in places

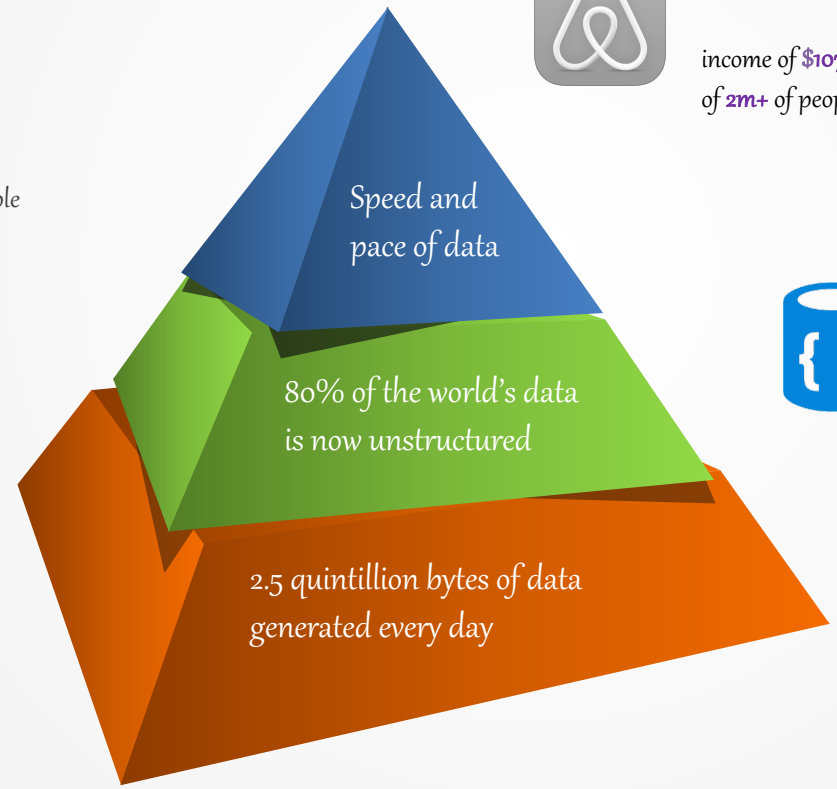
Lyft, Inc.

1m+ million of rides per day and 30M+ of users worldwide

Apache Kafka [2014]

speed of data generation & processing large datasets, ability to ingest data as fast as possible

Batch, Near, Real-Time



NoSQL [2009]

type & nature of data, different data sources & mappings, easier for developers

Key-Value Pairs, Column-Family, JSON, Graph

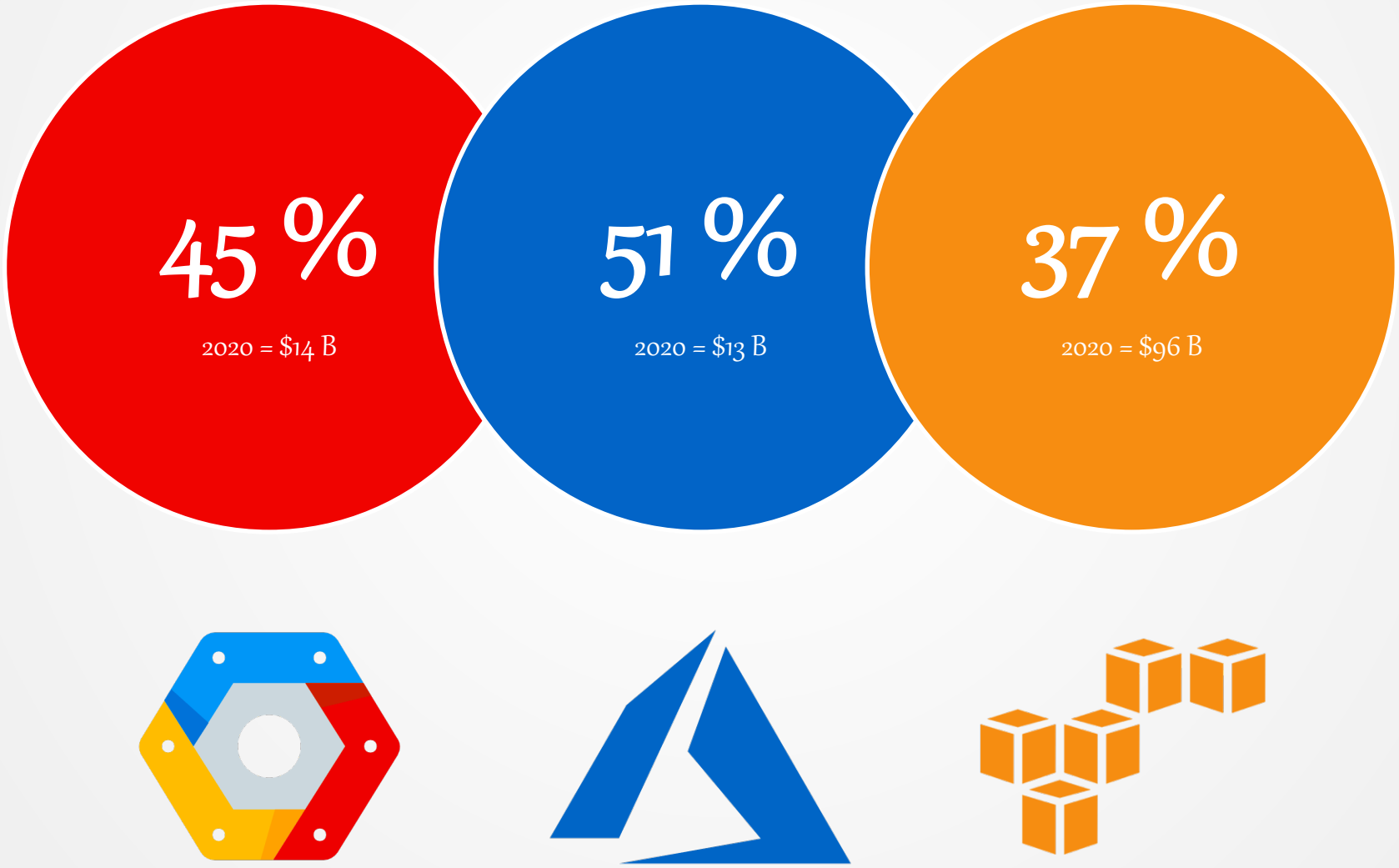
Apache Hadoop [2006]

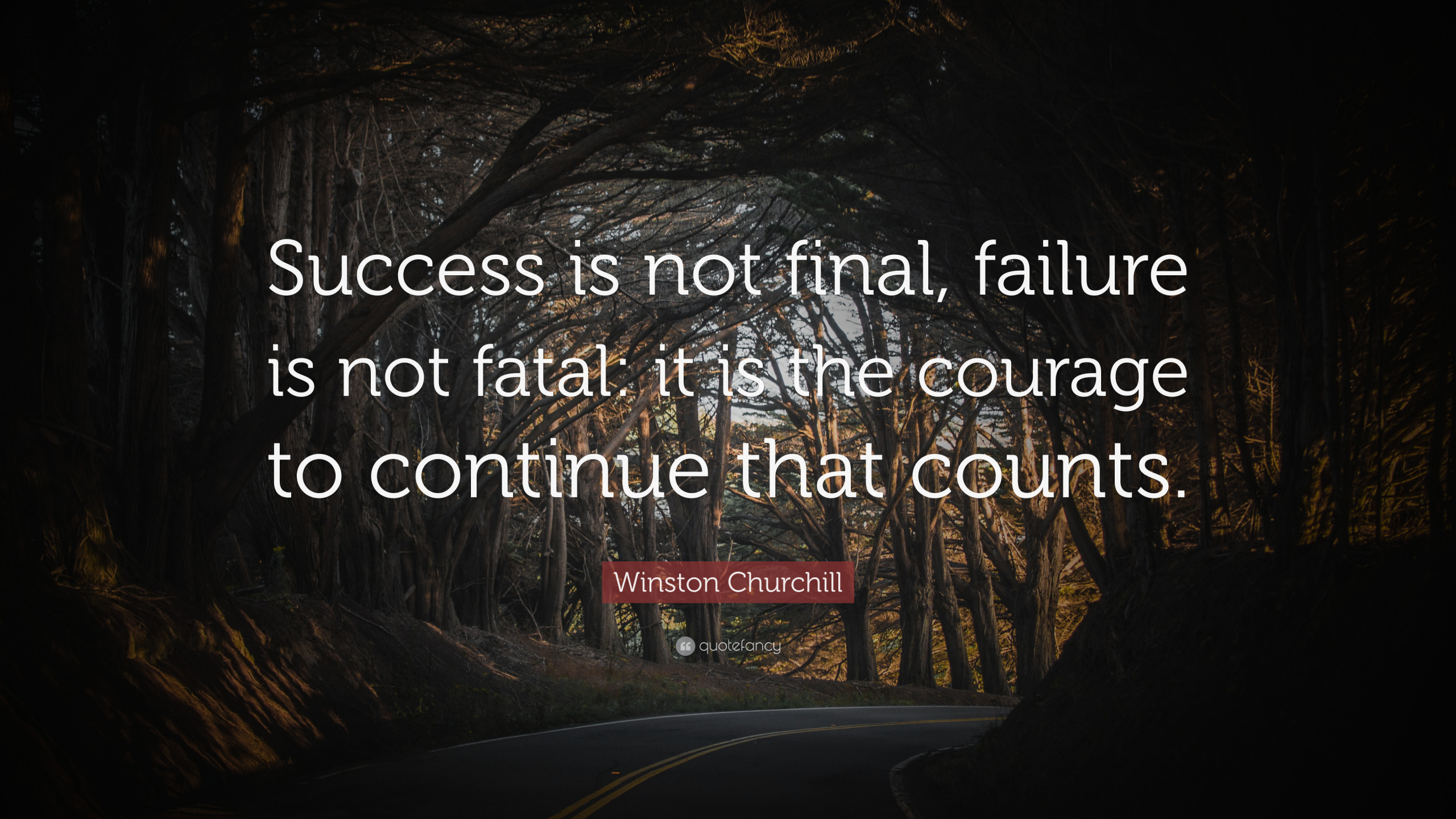
quantity of generated & stored data, size of data & value of potential insights

MB | GB | TB | PB

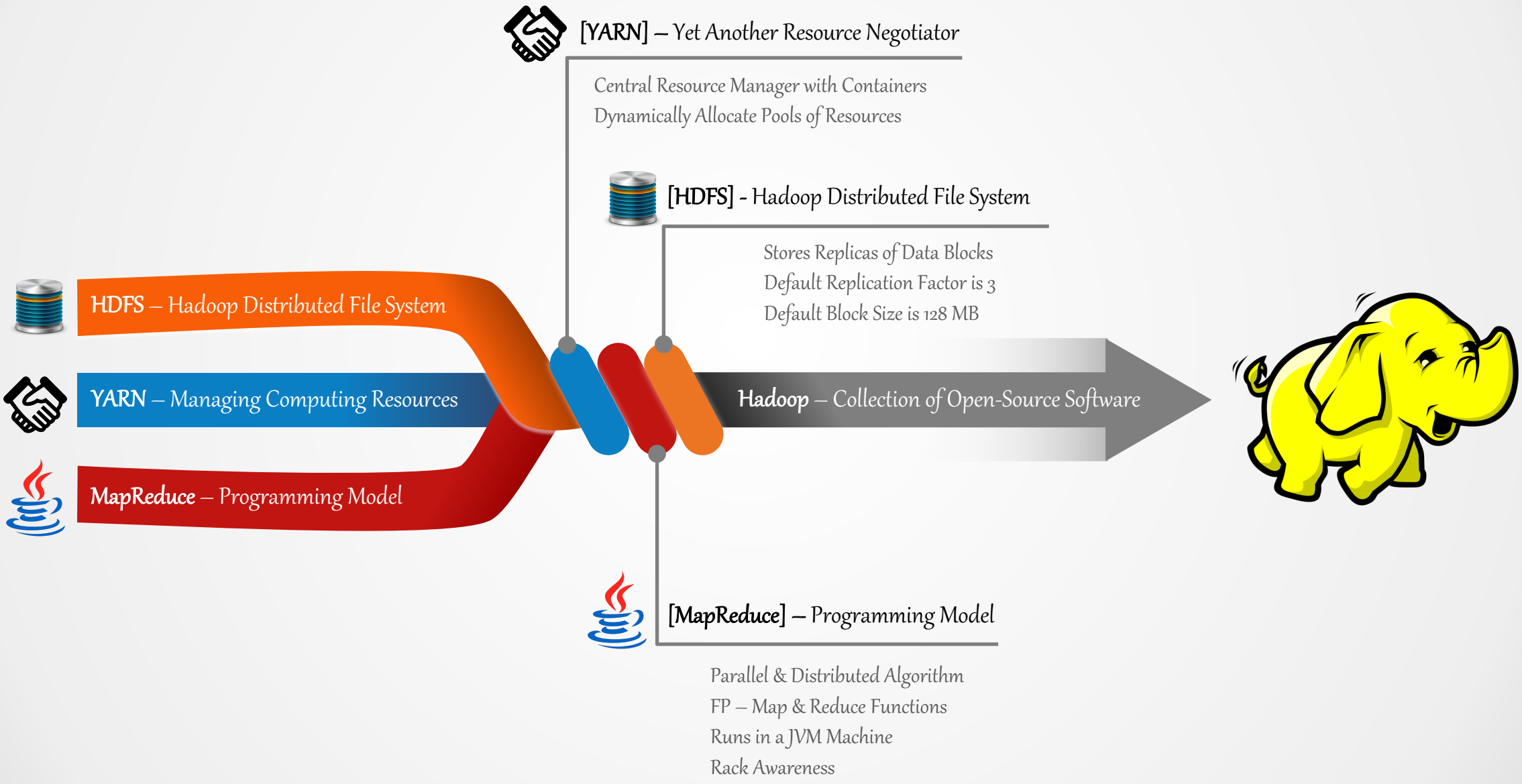

With enough courage, you can do without a reputation.
Margaret Mitchell
quotefancy

Google Cloud Platform vs. Microsoft Azure vs. Amazon Web Services




Success is not final, failure is not fatal: it is the courage to continue that counts.
Winston Churchill
quotefancy

Apache Hadoop [Fundamentals]



Ecosystem of Hadoop Animal Zoo [2006 ~ 2014]

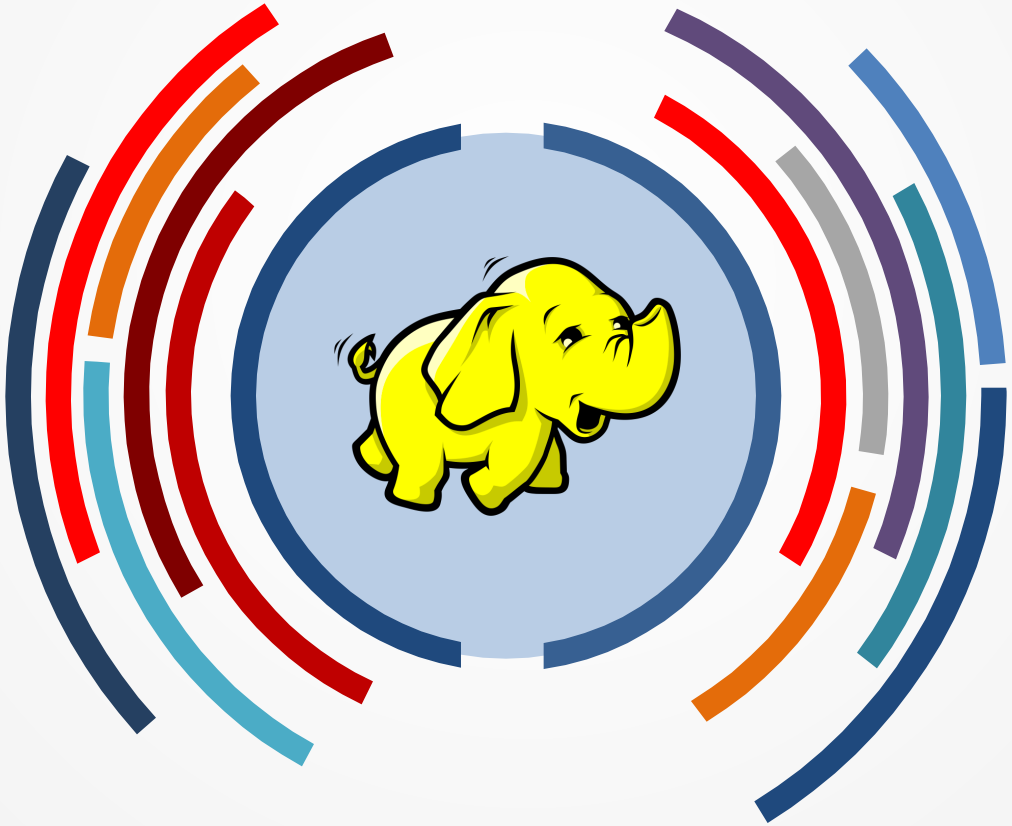
Apache Pig
High-Level Platform
Pig Latin for ETL Jobs

Apache Hive
Data Warehouse with SQL-Like Interface
Used By Facebook & Netflix

Apache HBase
Non-Relational Distributed Database
Used By Netflix & Spotify

Apache Phoenix
Massively Parallel & Relational Database
Skin of Apache HBase [ACID] | OLTP

Apache Zookeeper
Distributed Configuration Service
Sync & Name Registry



Apache Flume
Distributed & Reliable for Collecting &
Aggregating Large Amounts of Log Data

Apache Storm
Distributed Stream Processing
Computation & Acquired by Twitter

Apache Sqoop
Command-Line Interface for Transferring
Data Between Relational DB's & Hadoop

Apache Oozie
Server-Based Workflow Scheduling for
Hadoop Jobs

Apache Mahout
Scalable ML Focused with Collaborative
Filtering Clustering & Classification

History of Apache Hadoop & Apache Spark

Doug Cutting

Started Working on Nutch

Google

Published MapReduce Paper

Google

Published GFS Paper

Nutch

GFS & MapReduce Support

Hadoop

Cutting's Son's Yellow Plush Toy
Sorts 1.8 TB on 188 Nodes in 47.9 Hs
Yahoo Hadoop Cluster > 600 Machines

Hadoop

Yahoo Web Index with Hadoop
World Record – Fastest System Sort TB
Loading 10 TB a Day in Yahoo Clusters
Cloudera, Hadoop Distributor is Founded

Hadoop

Yahoo Runs 17 Clusters with 24,000 Machines
Sorts a PB of Data [1 PB in 62 Seconds]
HDFS & MapReduce as a SubProject
MapR, Hadoop Distributor Founded

2002 2003 2004 2005 2006 2008 2009

2020 2019 2018 2017 2014 2011 2010

Apache Spark

Spark 2.4.5 Released
Spark 2.4.6 Released
Spark 2.4.7 Released
Spark 3.0.0 Released
Spark 3.0.1 Released

Apache Spark

Spark 2.2.3 Released
Spark 2.3.3 Released
Spark 2.4.1 Released
Spark 2.4.2 Released
Spark 2.4.3 Released
Spark 2.4.4 Released

Hadoop

Apache Hadoop 3.1

Hadoop

Apache Hadoop 2.9
Apache Hadoop 3.0

Hadoop

Apache Hadoop 2.3
Apache Hadoop 2.4
Apache Hadoop 2.5
Apache Hadoop 2.6
Apache Spark Top Level ASF

Hadoop

Facebook, LinkedIN, eBay & IBM
200,000 Lines of Code
42K Hadoop Nodes
Top Prize at Media Guardian Innovation Awards
Rob Beardon & Eric Badleschieler Spin HortonWorks

Hadoop

Yahoo 4,000 Nodes & 70 PB
Facebook 2,300 Clusters & 40 PB
Apache HBase Graduates
Apache Hive Graduates
Apache Pig Graduates
Apache Zookeeper Graduates

Apache Spark

DataFrame & DataSet API Unified
Apache Spark 2.0 Released

Kappa Architecture

Jay Kreps
Principal Staff Engineer

Lambda Architecture

Nathan Marz
Software Engineer at Twitter

Big Data-as-a-Services [BDaaS]

Hadoop-as-a-Service [HaaS]

Collection of Open-Source Software

Fully-Managed Cloud Service
- Amazon AWS
- Microsoft Azure
- Google Cloud Platform

Amazon Elastic MapReduce [EMR]

Easy Run & Scale Big Data Frameworks

Apache Hadoop
Apache Spark
HBase
Presto & Hive

Cloud DataProc

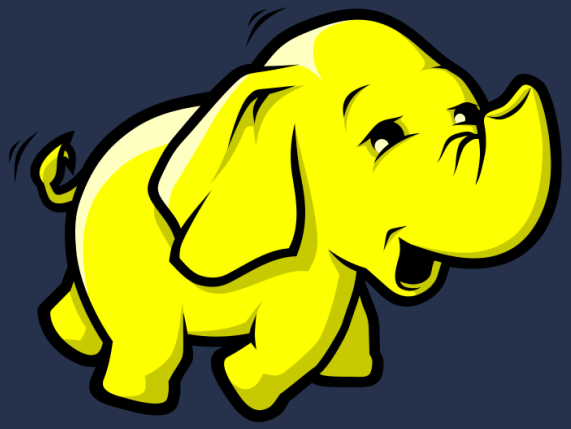
Fully-Managed Cloud Service
Cloud Native Apache Hadoop & Apache Spark
Provisioning Time of 90 Seconds

HDInsight

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

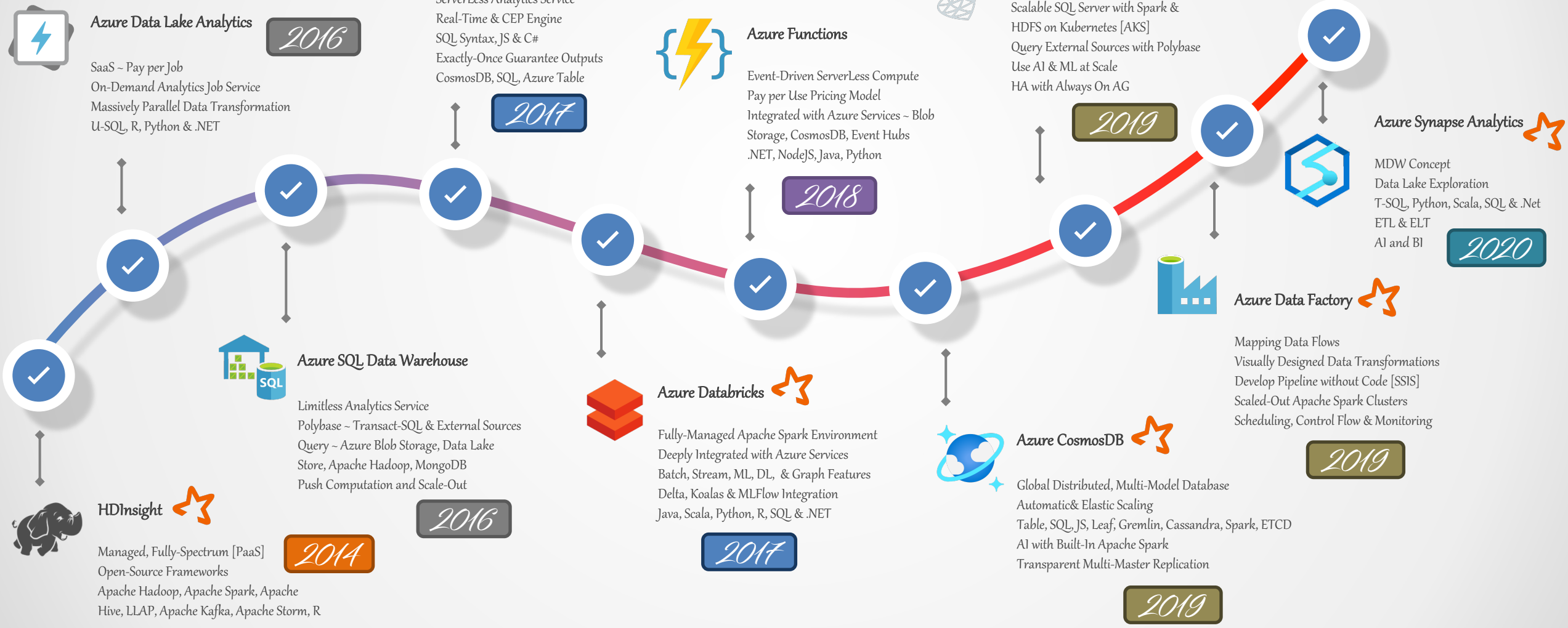
Apache Hadoop
Apache Spark
Apache Kafka
Apache HBase
Apache Hive LLAP
Apache Storm
Machine Learning

Develop Java MapReduce Program for Apache Hadoop [Locally]




Courage is never to let your actions be influenced by your fears.
Arthur Koestler
quotefancy

Data Processing Engines [TimeLine] ~ Microsoft Azure



Apache Spark [Fundamentals]

Apache Spark

Open-Source Distributed Cluster-Computing Framework
Implicit Data Parallelism & Fault Tolerance
Optimized for Memory Computation
Written In - Scala
100x ~ MapReduce Jobs & 10x – Disk-Based Operations

History

University of California, Berkeley's AMPLab
Open-Sourced in 2014 – Top-Level Apache Project
Databricks – New World Record in Large Scale Sorting [**2014**]
Ali Ghodsi | Reynold Xin | Matei Zaharia ~ **Databricks**
1,000 Contributors in 2015

Key Capabilities

Unified Stack for Interactive, Streaming & Predictive Analysis
Batch & Streaming in an Unified Platform
Designed for Large-Scale Data Processing

Performance

Daytona Gray

- 100 TB in **23 Minutes** with 206 EC2 VMs
- 100 TB in 72 Minutes with 2.100 Nodes

Use-Cases

AirBnB – ML & Streaming on Mobile App

Uber – Continuous ETL Pipeline by Processing Apache Kafka Data and Storing on HDFS

Netflix – Data Analytics Computation & ML at Scale

Core APIs

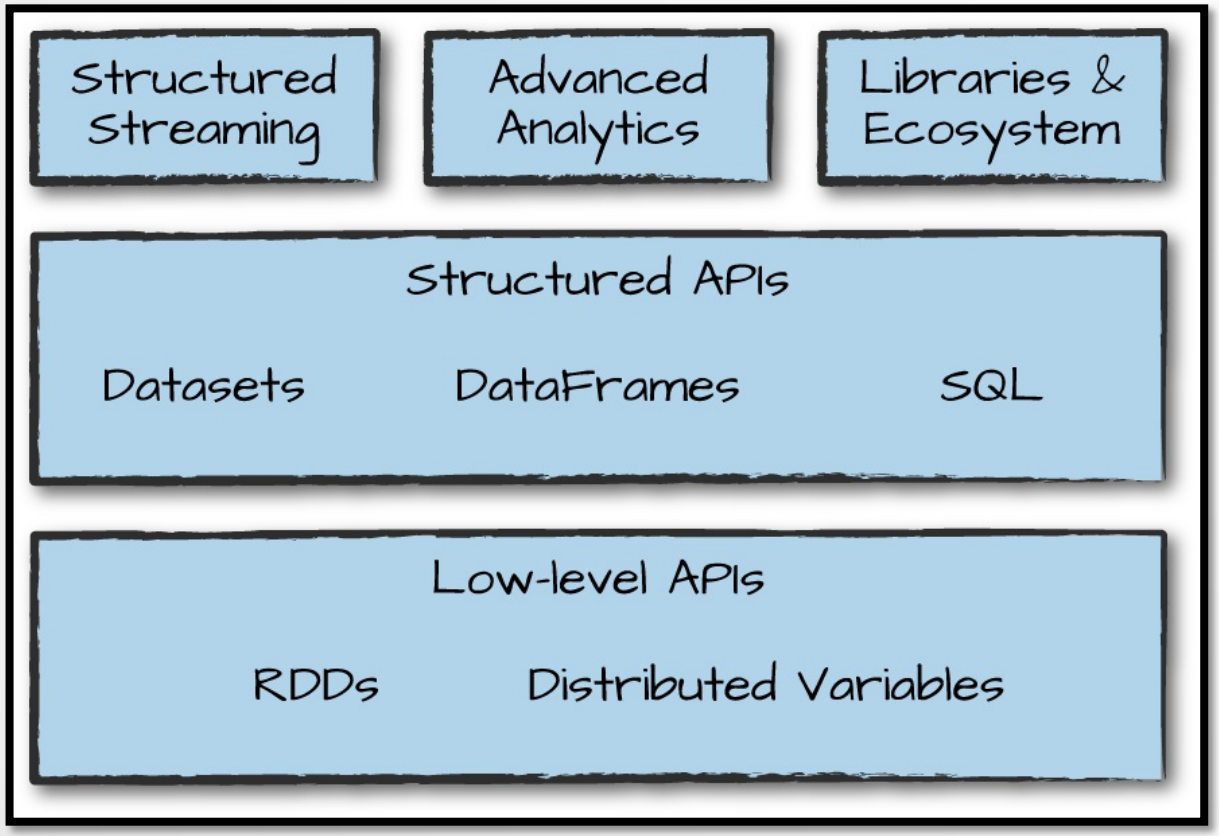
SQL
Java
Scala
Python
R

Processing Structures

RDD – Resilient Distributed DataSet
Spark Streaming – Processing Data Streams using DStreams
Spark-SQL, DataSets & DataFrames – Processing Structured Data
Structured Streaming – Processing Structured Data Streams

Apache Spark [Components]

developers stopped making individual processors faster and switched toward adding more parallel cpu cores all running at the same speed. this change meant that suddenly applications needed to be modified to add parallelism in order to run faster, which set the stage for new programming models such as apache spark.



High-Level APIs

the fundamental abstraction that you will use to write most of your data flows.

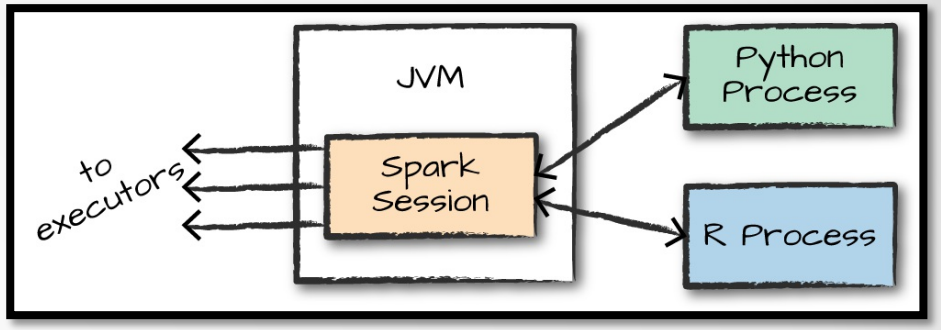
- fundamentally, spark is a distributed programming model in which the user specifies transformations, these transforms build a directed acyclic graph of instructions
- actions, begins the process of executing theses dags as a single job by breaking down into stages and tasks to execute across the cluster

Low-Level APIs

virtually everything in spark is built on top of RDDs.

- you need some functionality that you cannot find in higher level apis for example a very tight control over physical data placement across the cluster
- you need to maintain some legacy codebase written in RDDs
- add some custom shared variable manipulation

Apache Spark [Architecture]



Pandas UDF with Type Hints

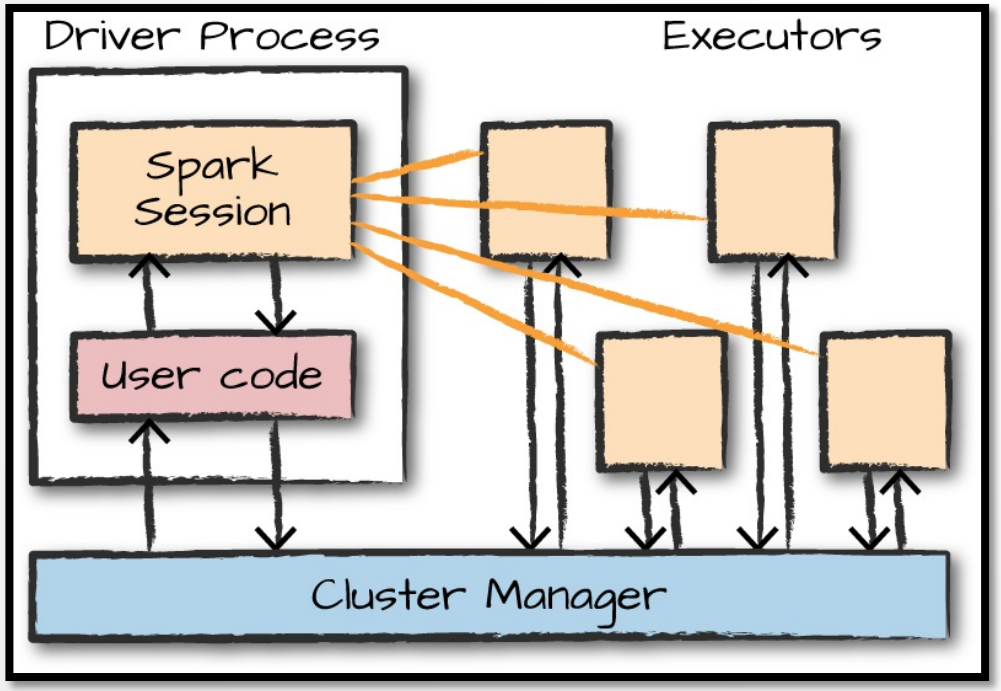
- Apache Arrow ~ JVM & Python Driver/Executor with Near Zero SerDes Cost
- Integration with Pandas API

Driver Process

the driver runs the main function, sits on a node in the cluster, and is responsible for

- maintaining information about the spark application
- responding to a user's program input
- analyze, distribute and schedule work across the executors

each language API maintains the same core concepts that we described earlier. there is a spark session object available to the user, which is the entrance point to running spark code. when using spark from python or r, you don't write explicit JVM instructions; instead, you write python and r code that spark translates into code that it then can run on the executor JVMs.



Executors

responsible for carrying out the work that the driver assigns them

- executing code assigned by the driver
- reporting the state of the computation

RDD | DataFrame & DataSet [Comparison]

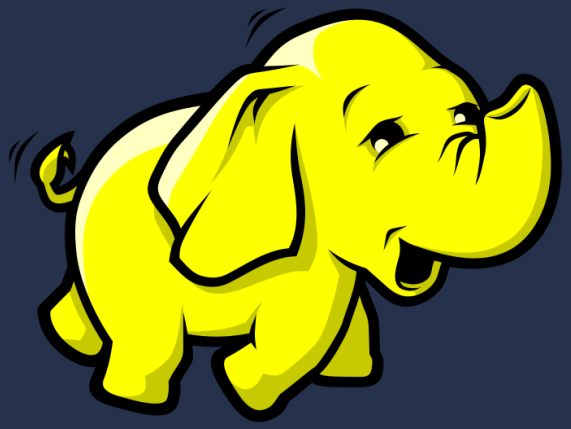
Apache Spark

DataFrame & DataSet API Unified

Apache Spark 2.0

	RDD	DataFrame	DataSet
Structured & Unstructured	✓	✓	✓
Java & Scala	✓	✓	✓
Python & R	✓	✓	✗
Any Data Source	✓	✗	✓
Schema Infer	✗	✓	✓
Optimization Engine	✗	✓	✓
Fast Aggregation	✗	✓	✓
In-Memory Serialization	✗	✓	✓

Develop PySpark Program using PyCharm [Local] & [YARN]



Apache Spark [Structured Streaming]

Philosophy [Moto]

Treat Data Streams = **Unbounded Tables**
Incremental Query over Streams

Guarantees

Scalable & Fault-Tolerant
Incremental & Continuous Data Streams
End-to-End Exactly-Once Guarantees
Checkpoint & Write Ahead Logs

Seamless

Express Streaming Computation = Batch Computation

Engine

Built on Spark SQL Engine
Project Tungsten [CPU & Memory]
Optimized using Catalyst Optimizer

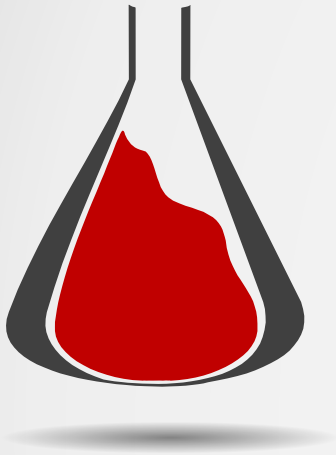
Processing

Uses Micro-Batch Processing Engine
100 milliseconds with Exactly-Once Latency

Spark ML [MLlib]

MLlib – Apache Spark's Scalable Machine Learning Library

RDD-Based API vs. DataFrame-Based API

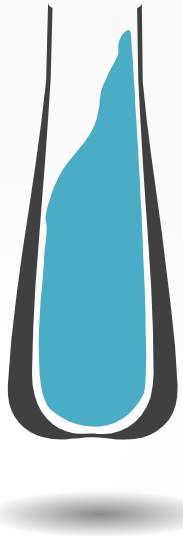


MLlib RDD-Based API [**spark.mllib**] Package [Maintenance Mode]
DataFrame-Based API [**spark.ml**] Package

DataFrame-Based API [Spark-ML]

- User-Friendly API
- Data Sources
- SQL & DataFrame Queries
- Tungsten & Catalyst Optimizations
- Facilitate Practical ML Pipelines [Feature Transformation]

ML Library [MLlib]



ML Algorithms – Classification | Regression | Clustering & Collaborative Filtering
Featurization – Extraction | Transformation | Dimensionality Reduction & Selection
Pipelines – Constructing | Evaluating & Tuning ML Pipelines
Persistence – Saving & Loading Algorithms, Models & Pipelines
Utilities – Linear Algebra | Statistics & Data Handling

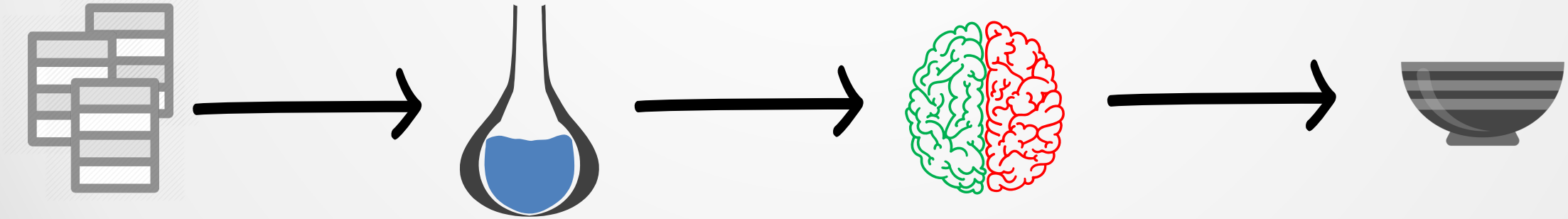
ML Pipeline

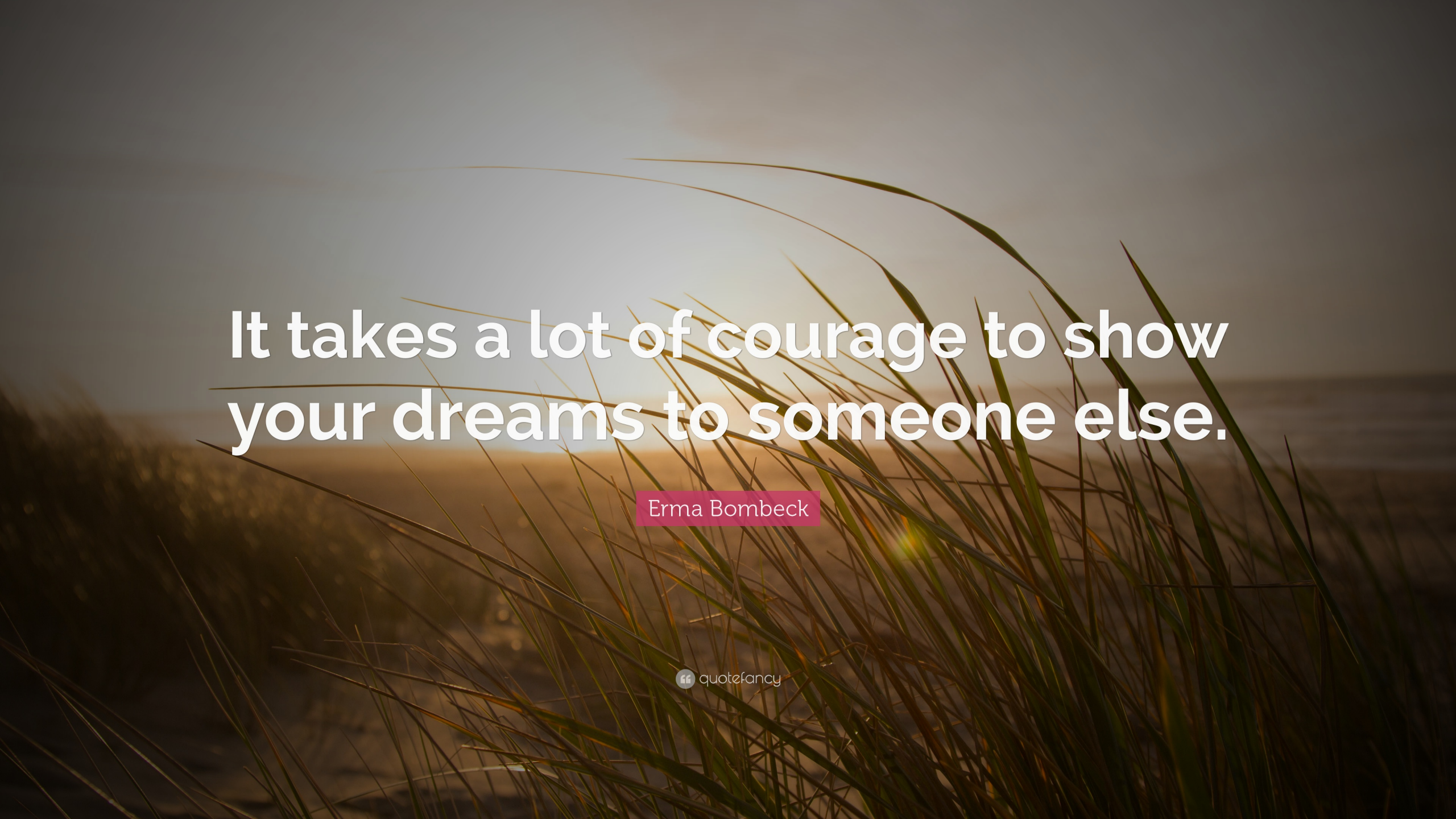
Dataset
Collection of Information

Feature Engineering
Domain Knowledge & Feature Creation

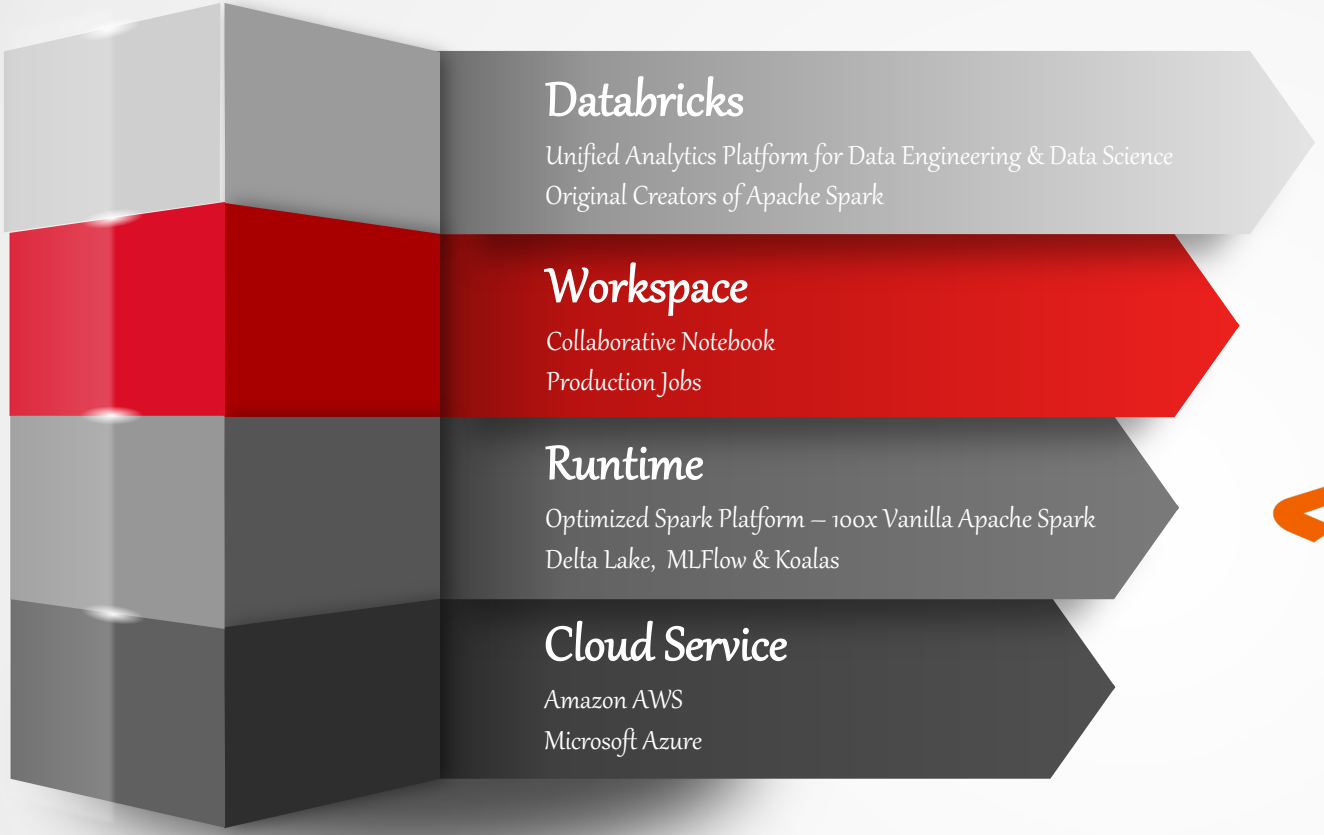
Model
Machine Learning Model [Algorithm]

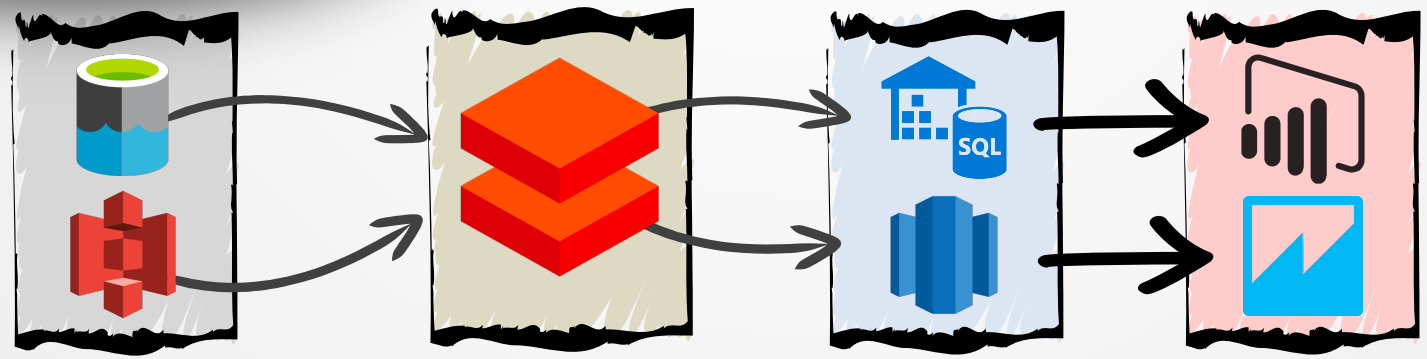
Serve
Serve Trained Model




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

Databricks [Overview]





Getting Started with Databricks

Creating Account on Databricks Community Edition


All things are difficult before they are easy.
Thomas Fuller
quotefancy

Lambda Architecture

Batch-Layer

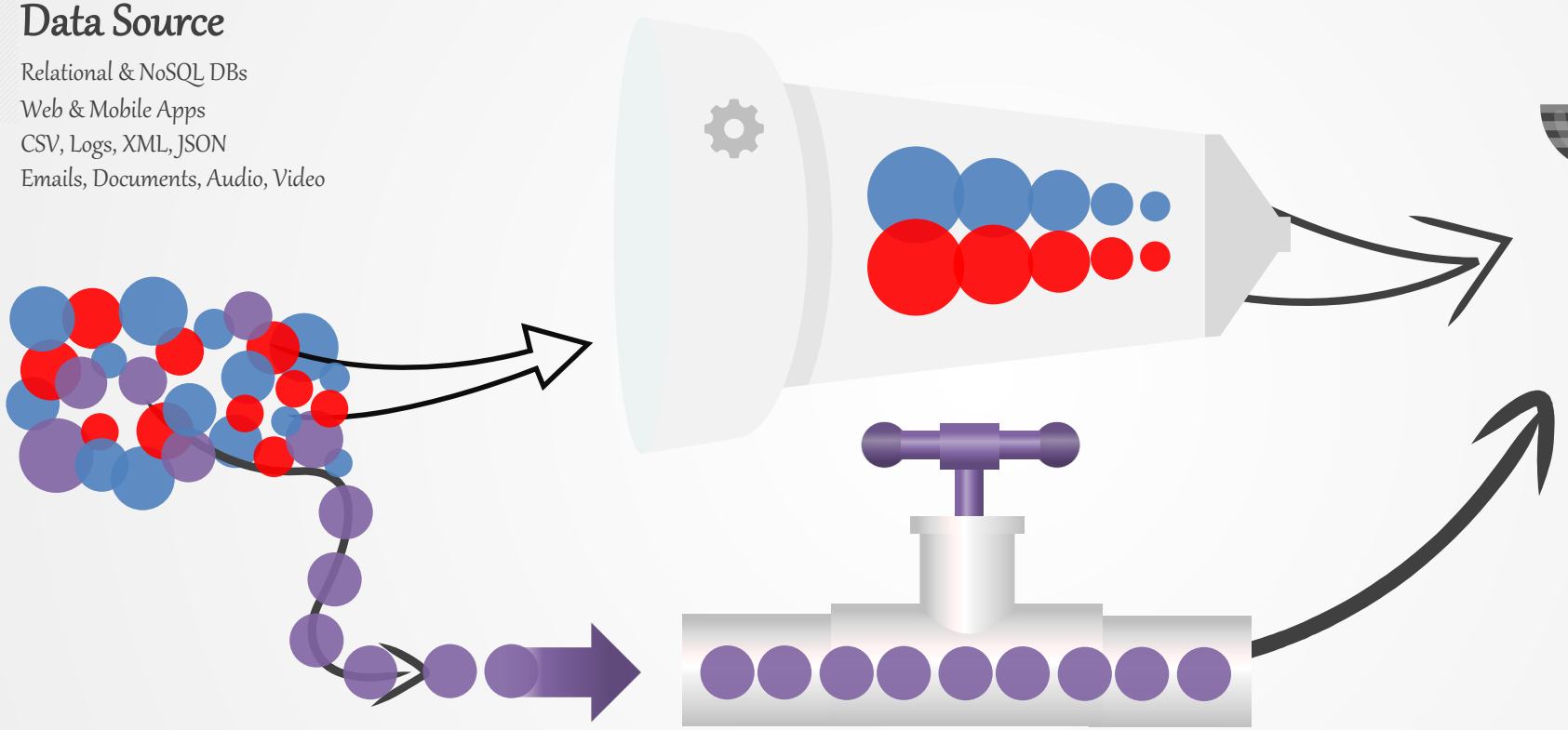
Data Storage - Data Stored in a Data Lake
Batch-Processing - New Data is Computed & Processed

Nathan Marz

Coined in 2011
Founder of Red Planet Labs
Software Engineer at Twitter

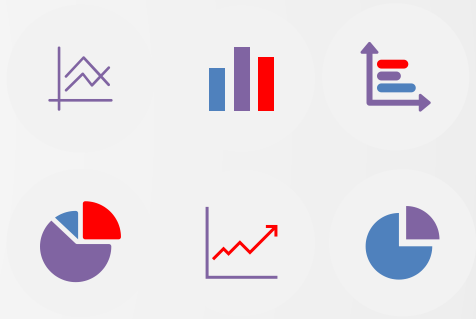
Data Source

Relational & NoSQL DBs
Web & Mobile Apps
CSV, Logs, XML, JSON
Emails, Documents, Audio, Video



Serving-Layer

Output of Batch & Speed Layer
Processed & Computed Data
Ad-Hoc Queries & Data Analysis



Speed-Layer

Real-Time Ingestion - Feed & Store Data in Real-Time
Stream Processing - Compute & Transform Inbound Data

Lambda Architecture – Cloud Agnostic & Simplified

Data Source

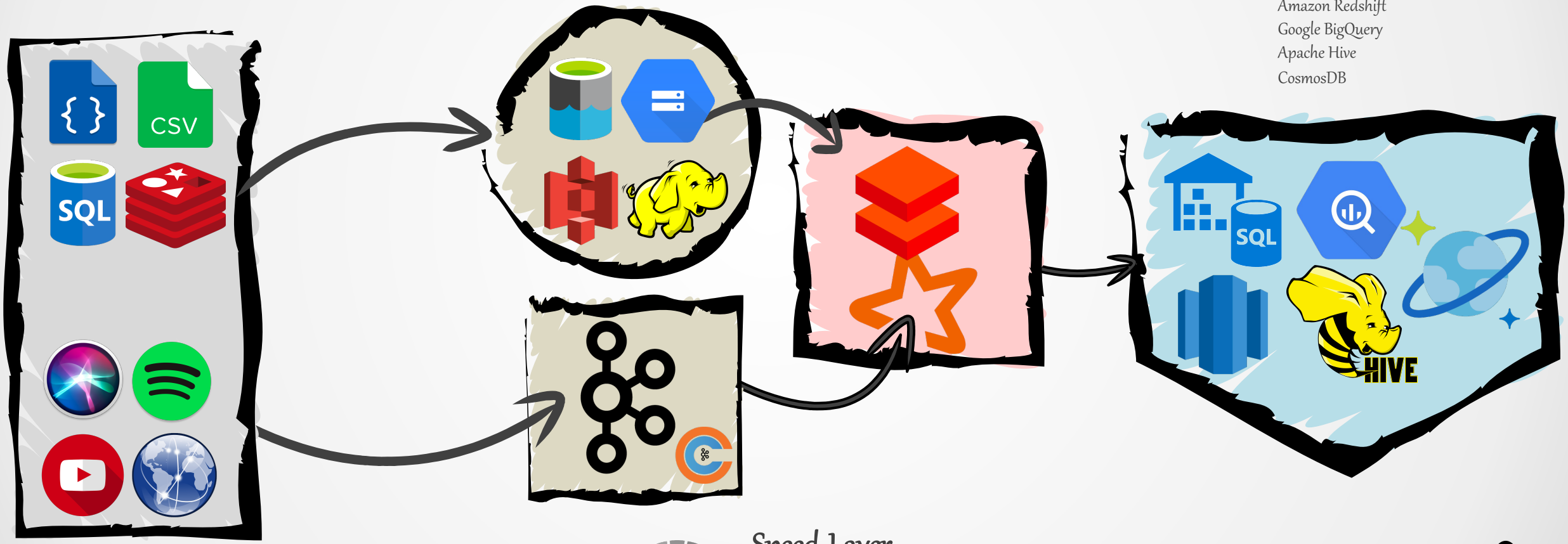
JSON & CSV
SQL Server & Redis
Internet – Siri | Spotify | YouTube

Batch-Layer

Data Storage - Data Lake Storage Gen2 | GCS | S3 | HDFS
Batch-Processing - Apache Spark | Databricks

Serving-Layer

Azure SQL Data Warehouse
Amazon Redshift
Google BigQuery
Apache Hive
CosmosDB



Speed-Layer

Real-Time Ingestion - Apache Kafka [Confluent]
Stream Processing - Apache Kakfa [Confluent] | Apache Spark [Databricks]

Kappa Architecture

Jay Kreps

Coined in 2014
Co-Founder & CEO at Confluent
Principal Staff Engineer

Data Source

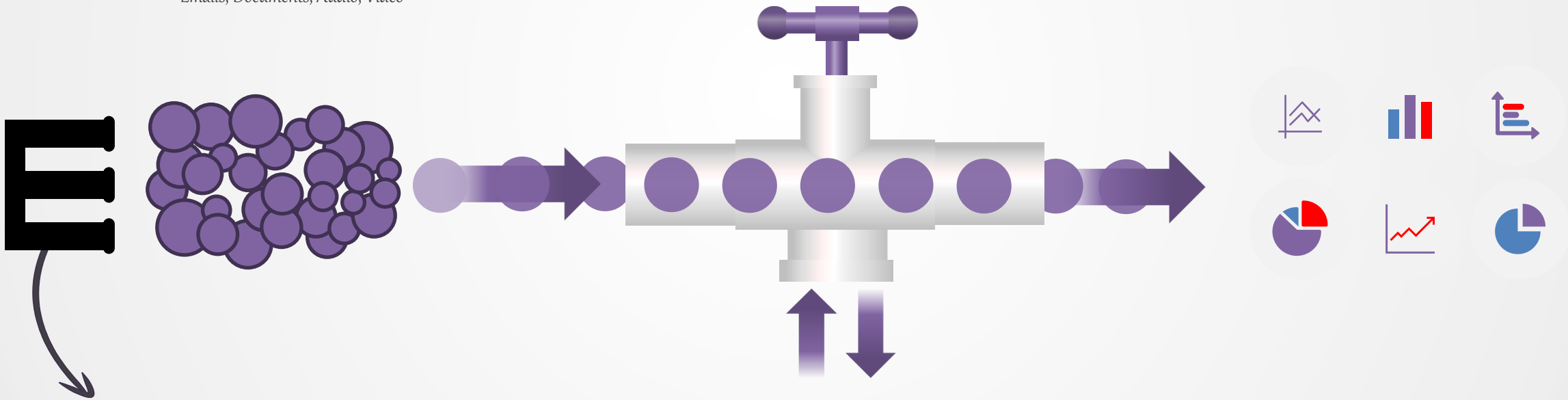
Relational & NoSQL DBs
Web & Mobile Apps
CSV, Logs, XML, JSON
Emails, Documents, Audio, Video

Speed-Layer

Real-Time Ingestion - Feed & Store Data in Real-Time
Stream Processing - Compute & Transform Inbound Data

Serving-Layer

Output of Batch & Speed Layer
Processed & Computed Data
Ad-Hoc Queries & Data Analysis



Event-Source

Unified Log Entry [Event Data]
Append-Only Log Store

Long Term Store

Keeps Data Stored for a Long Period of Time
Backfilling Process Capability

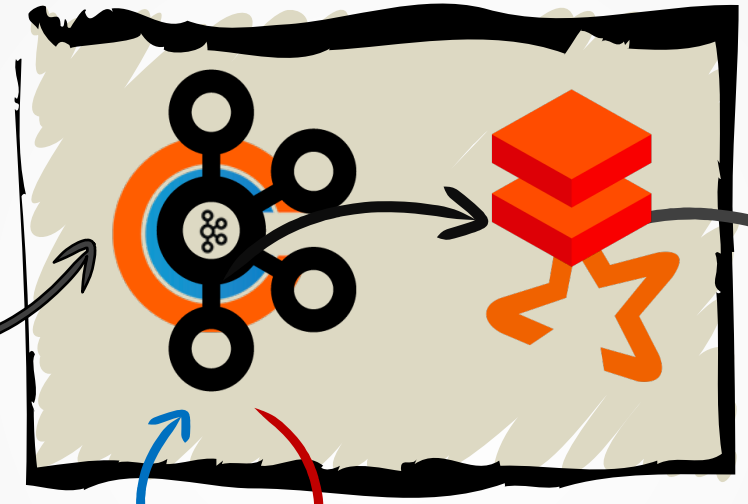
Kappa Architecture – Cloud Agnostic & Simplified

Speed-Layer

Apache Kafka ~ [Confluent]
Apache Spark ~ [Databricks]

Serving-Layer

Output of Batch & Speed Layer
Processed & Computed Data
Ad-Hoc Queries & Data Analysis

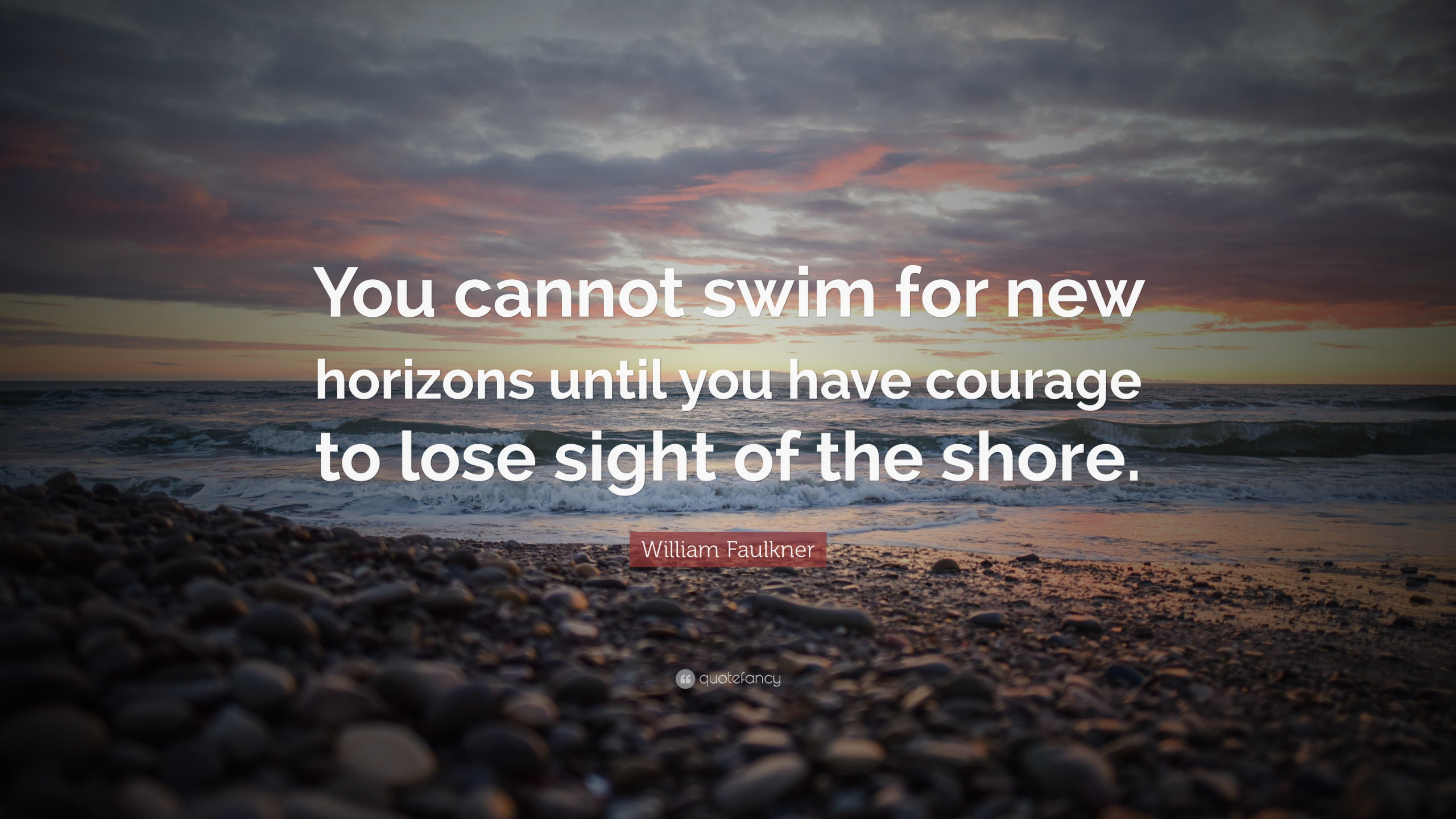


Event-Source

Unified Log Entry [Event Data]
Append-Only Log Store

Long Term Store

Azure Data Lake Storage Gen2
Google Cloud Storage [GCS]
Amazon S3
Hadoop Distributed File System [HDFS]


You cannot swim for new horizons until you have courage to lose sight of the shore.
William Faulkner
quotefancy

Use Case ~ Batch-ETL

Data Lake

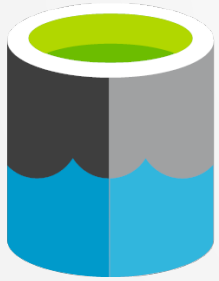
Repository of Raw Data
Without Schema Enforcement

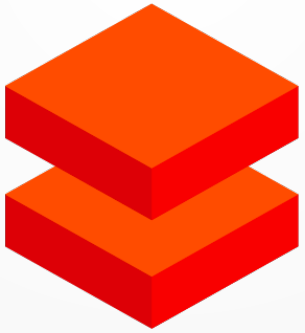
Apache Spark

Distributed Cluster-Computing Framework
Optimized for Memory Computation

Delta Lake

Storage Layer with ACID Transactions ~
Apache Spark & Big Data Workloads





ONEWAY

SOLUTION