ONEWAY

SOLUTION

One Way Solution

Foundation of Apache Spark

Data Engineering – [Day1]

LUAN MORENO
CEO & CDO
Data Engineer & Data Platform MVP


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

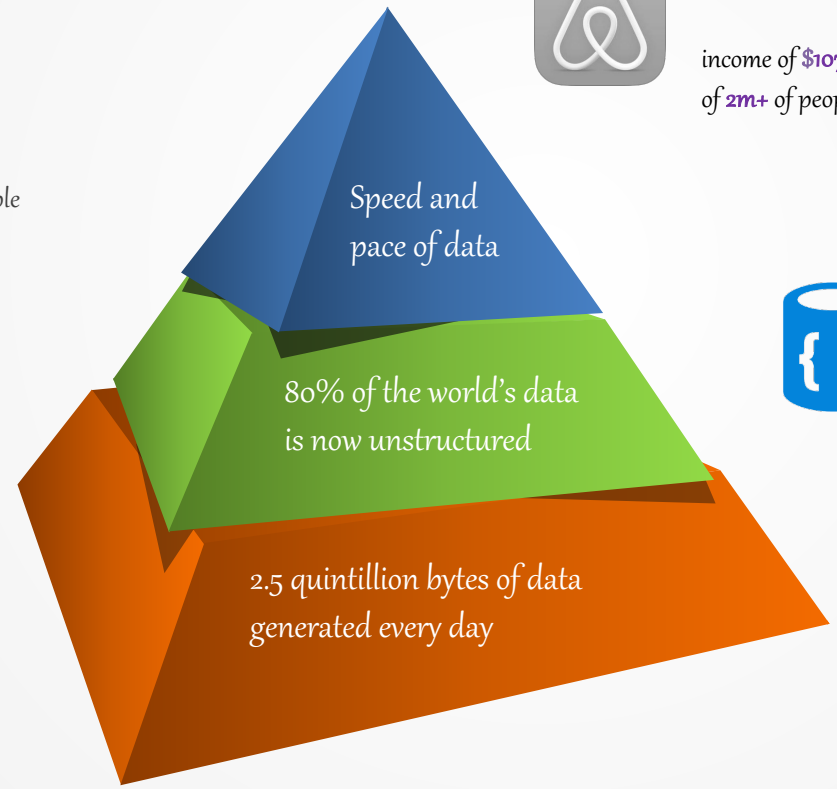
Lyft, Inc.

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]

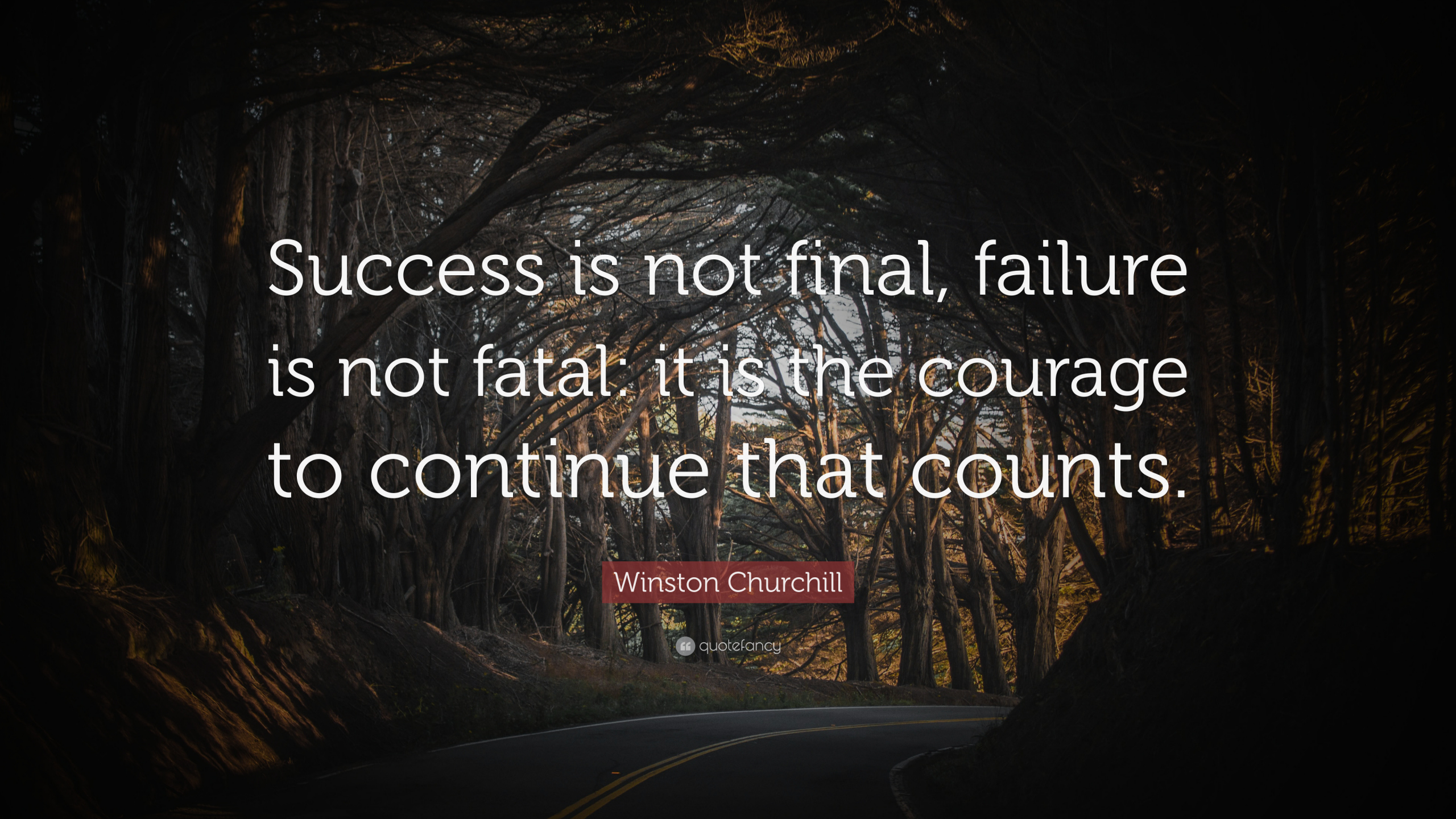
type & nature of data, different data sources & mappings, easier for developers

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

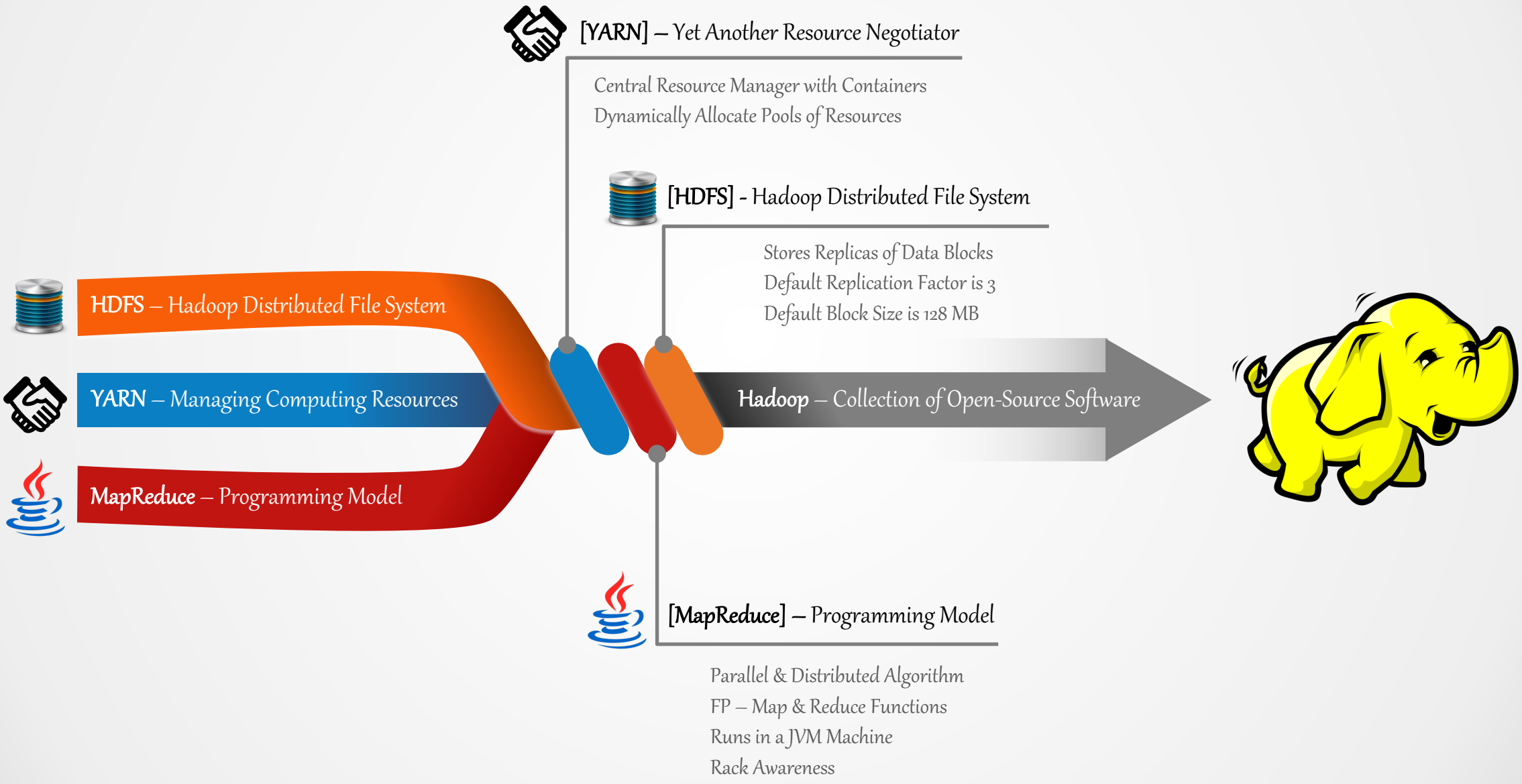
Apache Hadoop [2006]

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

MB | GB | TB | PB


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]

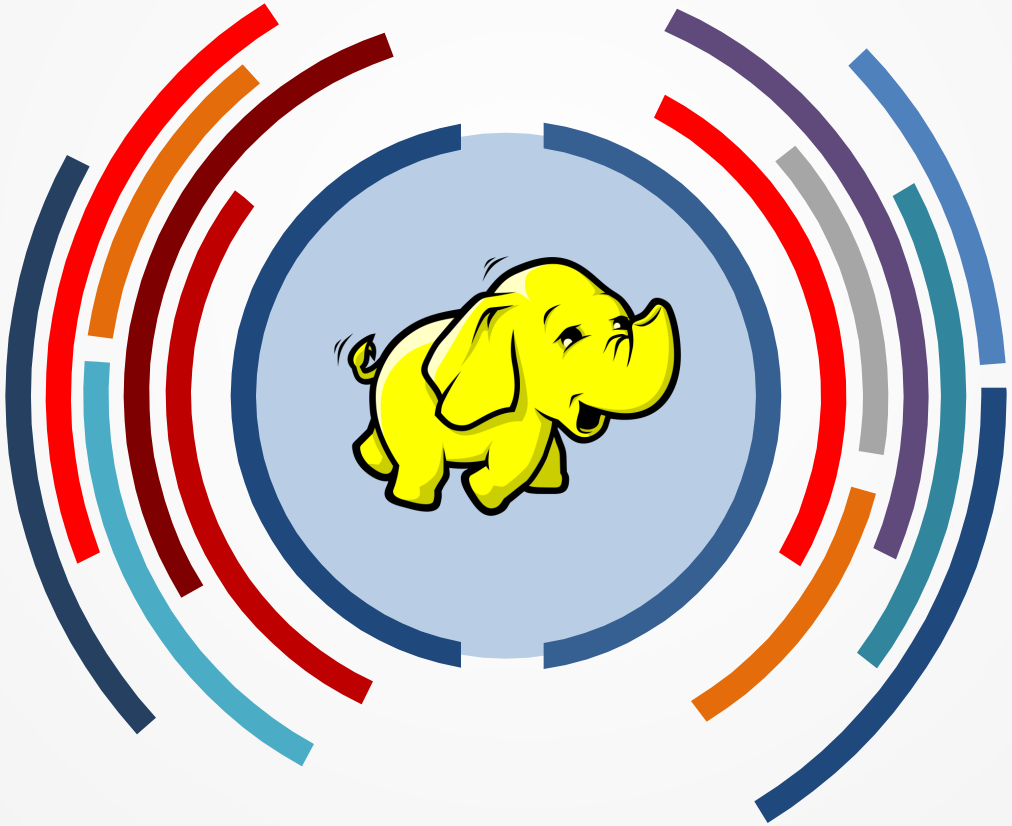
Apache Pig
High-Level Platform
Pig Latin for ETL Jobs

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

2022 2021 2020 2019 2018 2017 2014 2011 2010

Apache Spark

Spark 3.1.3 Released
Spark 3.2.1 Released

Apache Spark

Spark 3.0.2 Released
Spark 3.1.1 Released
Spark 3.1.2 Released

Apache Spark

Spark 2.4.5 Released
Spark 2.4.6 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

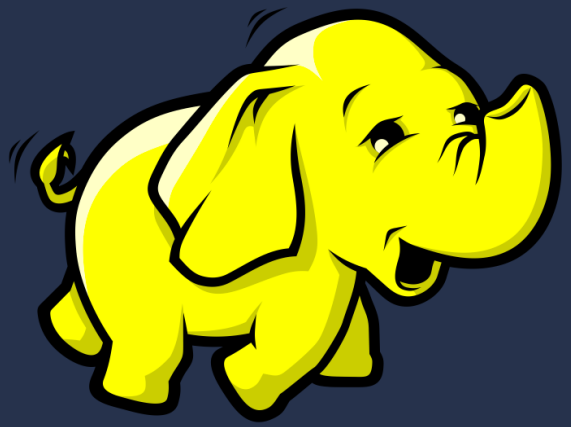
Fully-Managed Cloud Service
Cloud Native Apache Hadoop & Apache Spark
Provisioning Time of 90 Seconds

HDInsight

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

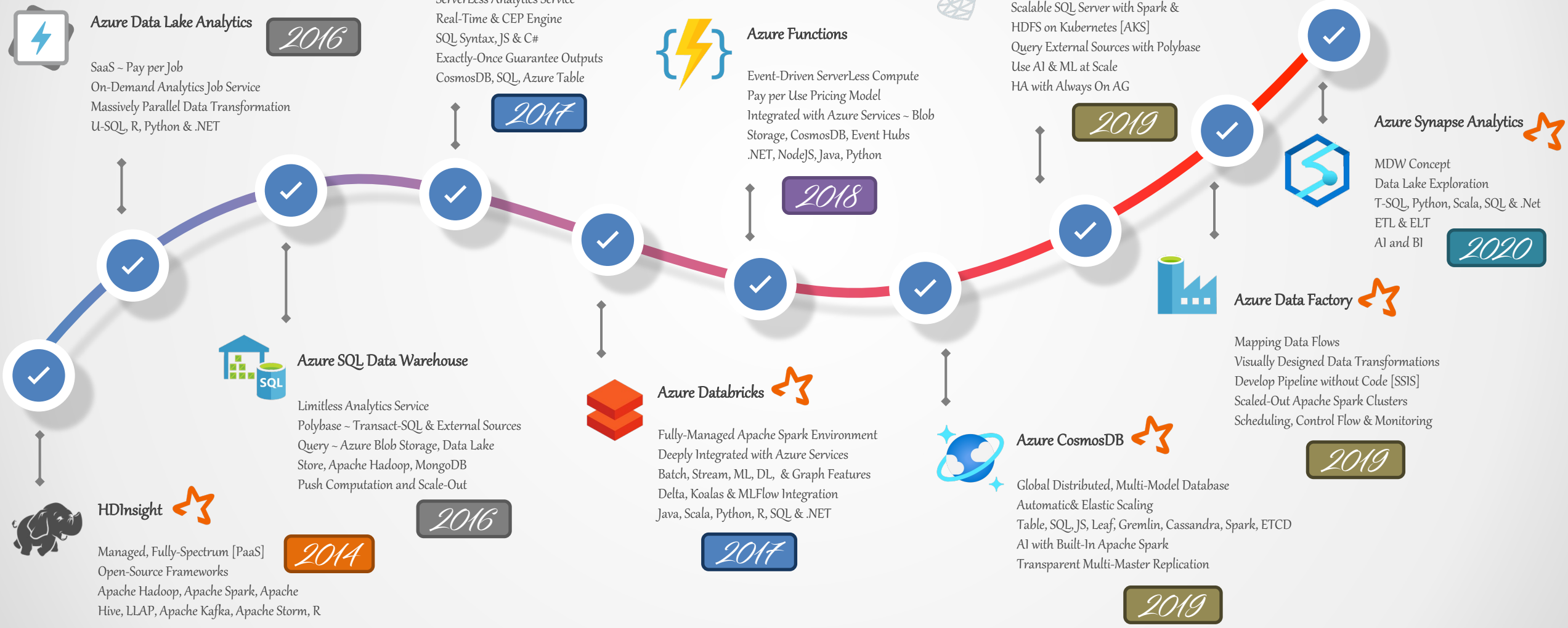
Apache Hadoop
Apache Spark
Apache Kafka
Apache HBase
Apache Hive LLAP
Apache Storm
Machine Learning

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



HDInsight

2014

Managed, Fully-Spectrum [PaaS]
Open-Source Frameworks
Apache Hadoop, Apache Spark, Apache Hive, LLAP, Apache Kafka, Apache Storm, R

Azure Data Lake Analytics

2016

SaaS ~ Pay per Job
On-Demand Analytics Job Service
Massively Parallel Data Transformation
U-SQL, R, Python & .NET

Azure SQL Data Warehouse

2016

Limitless Analytics Service
Polybase ~ Transact-SQL & External Sources
Query ~ Azure Blob Storage, Data Lake Store, Apache Hadoop, MongoDB
Push Computation and Scale-Out

Azure Stream Analytics

2017

ServerLess Analytics Service
Real-Time & CEP Engine
SQL Syntax, JS & C#
Exactly-Once Guarantee Outputs
CosmosDB, SQL, Azure Table

Azure Databricks

2017

Fully-Managed Apache Spark Environment
Deeply Integrated with Azure Services
Batch, Stream, ML, DL, & Graph Features
Delta, Koalas & MLFlow Integration
Java, Scala, Python, R, SQL & .NET

Azure Functions

2018

Event-Driven ServerLess Compute
Pay per Use Pricing Model
Integrated with Azure Services ~ Blob Storage, CosmosDB, Event Hubs
.NET, NodeJS, Java, Python

Azure CosmosDB

Global Distributed, Multi-Model Database
Automatic& Elastic Scaling
Table, SQL, JS, Leaf, Gremlin, Cassandra, Spark, ETCD
AI with Built-In Apache Spark
Transparent Multi-Master Replication

2019

SQL Server Big Data Clusters

Scalable SQL Server with Spark & HDFS on Kubernetes [AKS]
Query External Sources with Polybase
Use AI & ML at Scale
HA with Always On AG

2019

Azure Data Factory

Mapping Data Flows
Visually Designed Data Transformations
Develop Pipeline without Code [SSIS]
Scaled-Out Apache Spark Clusters
Scheduling, Control Flow & Monitoring

2019

Azure Synapse Analytics

MDW Concept
Data Lake Exploration
T-SQL, Python, Scala, SQL & .Net
ETL & ELT
AI and BI

2020

Power Bi

2020

Data Flows ~ Advanced Data Preparation
Collection of Data
Connect with Business from Sources,
Clean, Transform and Load into Storage

Apache Spark on [Cloud Providers] as a Product

Azure Data Factory Mapping Data Flows

Visually Designed Data Transformations in Azure Data Factory. Engineers to Develop and Write Logic without Writing Code. Use Scale-Out Apache Spark Clusters.

Integration with Azure Services for Better Experience. Offers Latest Data Processing Transformations.

Monitoring, Lineage and Metadata Visualization

AWS Glue

Serverless Data Integration Service for Discover, Prepare and Combine Data for Analytics, ML & Application Development. Run Python & Scala Code using Apache Spark Engine

Provides Visual & Code-Based Interfaces for Data Integration Workflows. Use AWS Glue Data Catalog for Data Governance and Metadata.

AWS Glue = Data Engineers
AWS Glue DataBrew = Data Scientists
AWS Glue Elastic Views = SQL

Cloud Data Fusion

Fully-Managed, Cloud Native Data Integration at Any Scale. Visual Point Interface for Code-Free Enablement, ETL & ELT Data Pipelines.

+150 Pre-Configured Connectors and Transformations
Natively Integrated Best-In-Class Google Cloud Services
End-to-End Data Lineage and RCA
Built with Open-Source CDAP for Pipeline Portability

Uses Apache Spark Cluster for Data Processing ~ Google Cloud Data Proc

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

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]

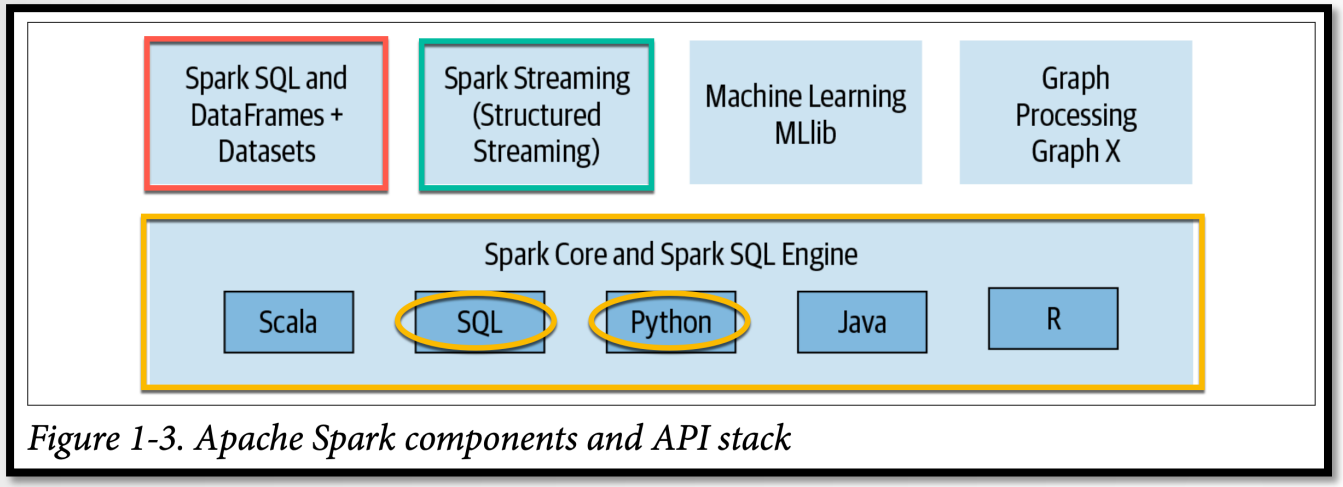


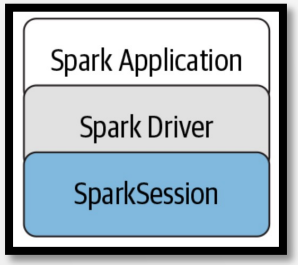
Figure 1-3. Apache Spark components and API stack

SQL & DataFrames + DataSet

- Read Structured Data
- Formats ~ CSV, Text, JSON, Avro, ORC, Parquet, Delta, Iceberg, Hudi
- Permanent or Temporary Tables In-Memory
- Use SQL Interface for Data Accessing

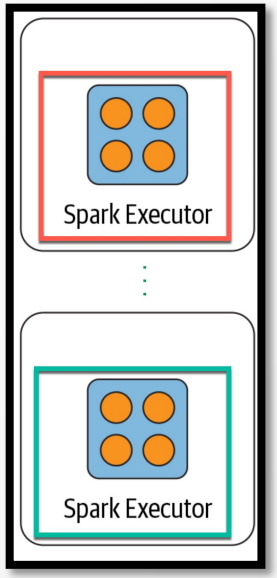
Structured Streaming

- Continuous Streaming Model using Structured Streaming API
- Uses SQL Engine & DataFrames-Based
- Lightweight, Faster and Better Integration



Responsible for Instantiating a Session, Communicates with Cluster Manager, Request Resources, Transforms DAG Computation, Schedule and Distribute Tasks.

Responsible for Managing and Allocating Resources for Cluster of Nodes.



Runs on Worker Node on Cluster. Communicate with Driver Program and Responsible for Executing Tasks.

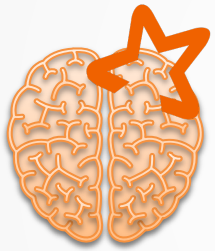
Apache Spark [Deployment Modes]

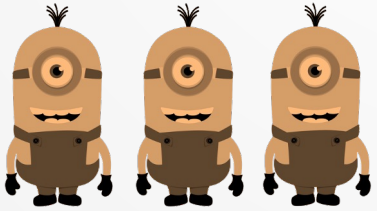
Local

Runs on a Single JVM ~ Laptop or Single Node, Executor and Cluster Manager Runs on Same Host

Standalone

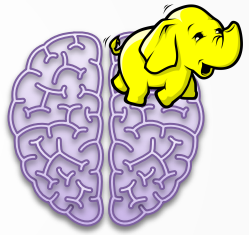
Run on Any Node in Cluster. Each Node Launch Own Executor JVM. Allocated Arbitrarily to Any Host

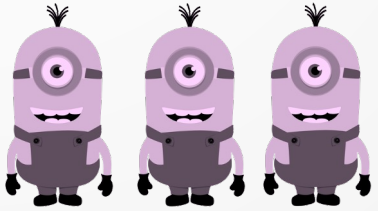




Yarn

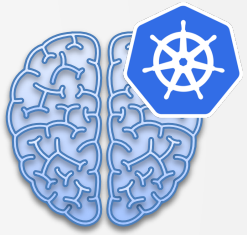
Runs with YARN (Cluster) Application Master. Node Manager & Resource Allocate Containers for Execution

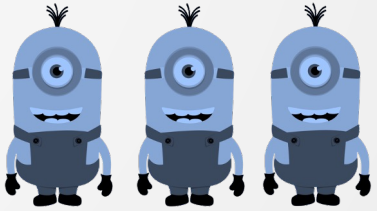




Kubernetes

Runs in a Kubernetes Pod. Each Worker Runs Within on Pod Context, use Kubernetes Master for Cluster Management





Apache Spark [Distributed Data & Partitions]

Data Lake

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

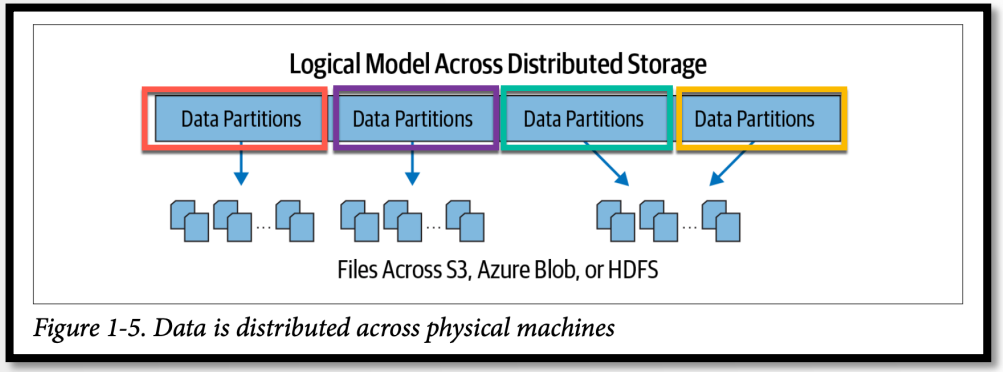
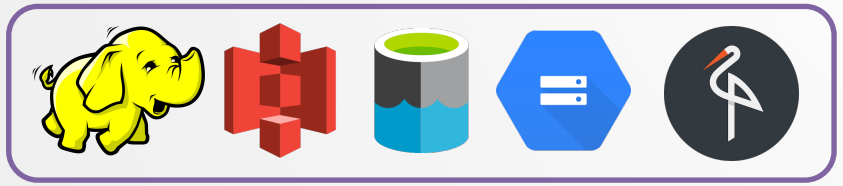
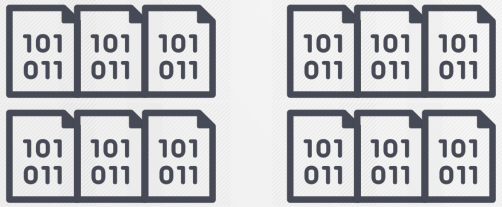


Figure 1-5. Data is distributed across physical machines





Processing Memory Engine

DataFrames In-Memory, Network Location, Data Locality.
Partitioning for Efficient Parallelism

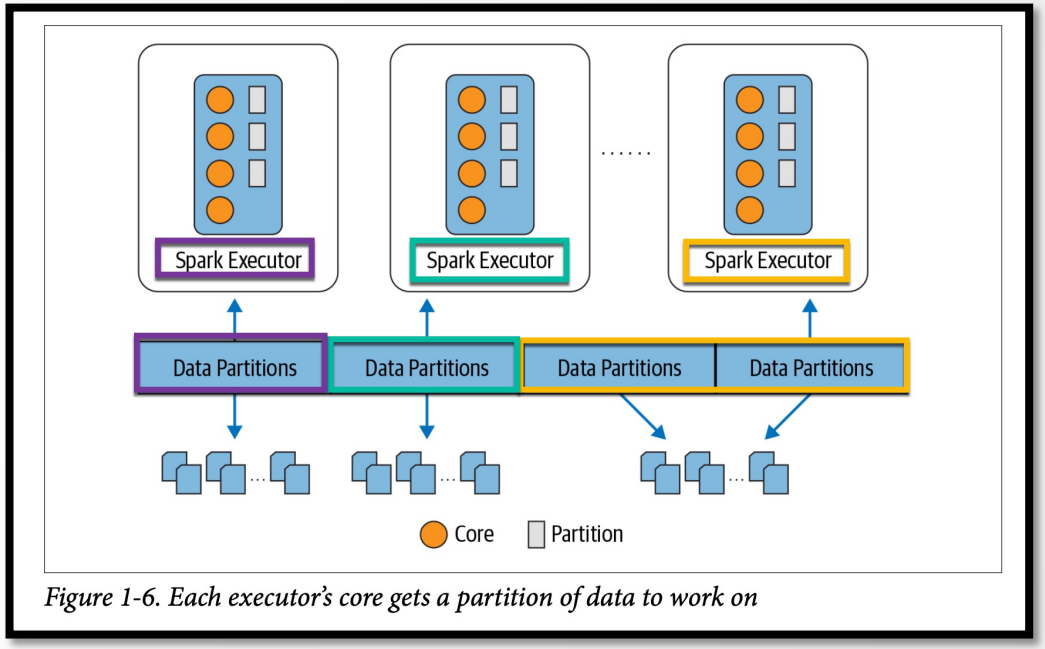
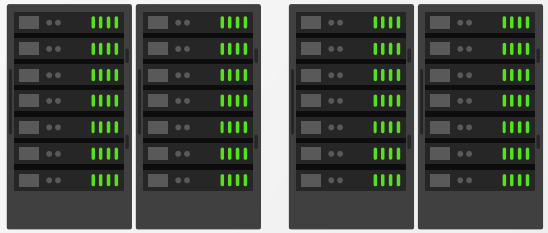


Figure 1-6. Each executor's core gets a partition of data to work on



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	✗	✓	✓

Dataframes vs. Datasets vs. RDDs

DataFrames

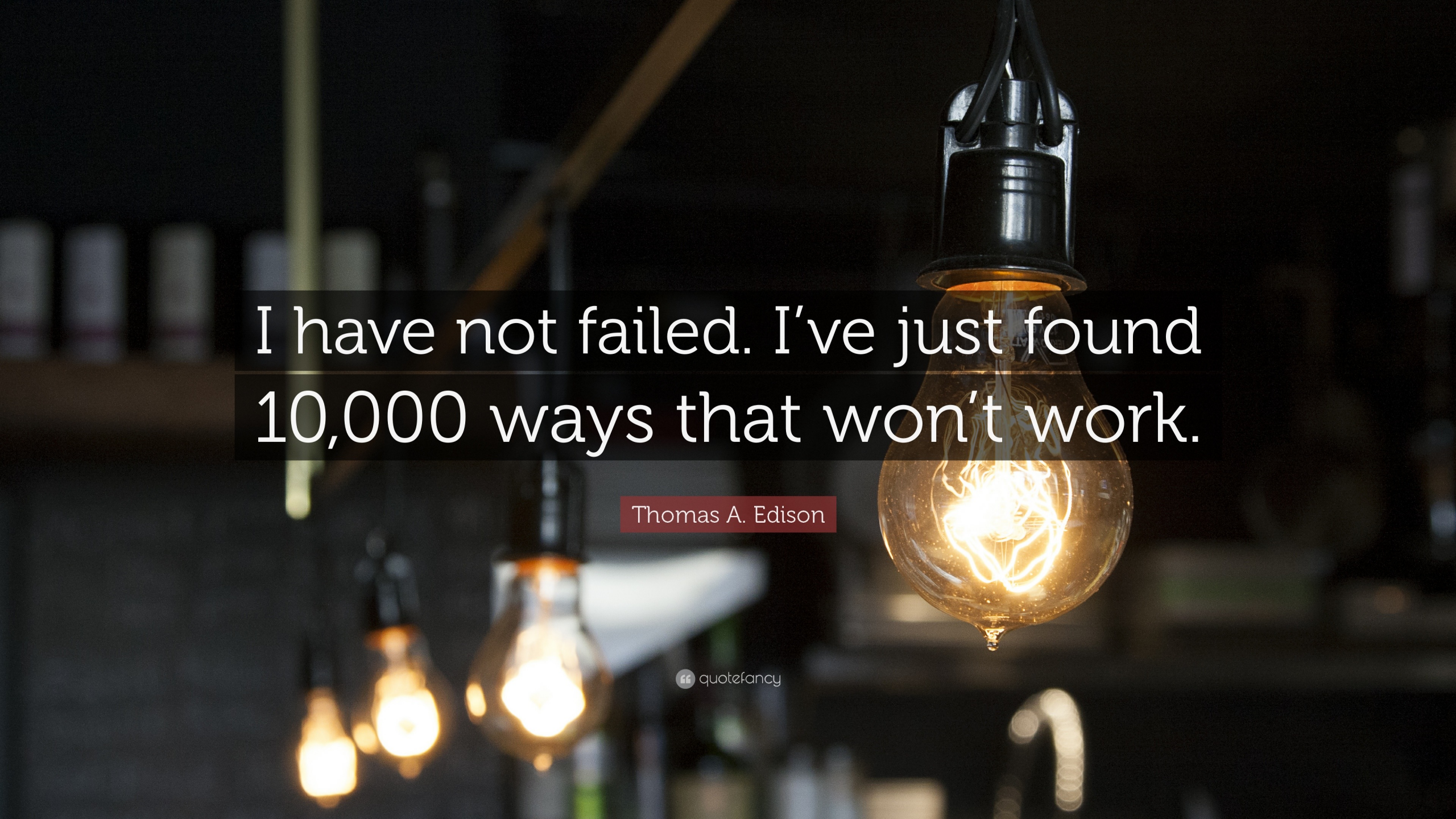
- Best Choice in Most Situations ~ 95%
- Provides Query Optimization using Catalyst
- Whole-Stage Code Generation
- Direct Memory Access
- Low GC Overhead
- For Extensibility Not as Developer-Friendly as DataSets

DataSets

- Good in Complex ETL Pipelines
- Not Good in Aggregations
- Provides Query Optimization using Catalyst
- Developer-Friendly by Providing Domain Object Programming & Compile-Time Checks

Resilient Distributed DataSet [RDD]

- Not Recommended Approach
- Without Query Optimization using Catalyst
- Without Whole-Stage Code Generation
- High GC Overhead
- For Apache Spark 1.X Legacy APIs


I have not failed. I've just found 10,000 ways that won't work.
Thomas A. Edison
quotefancy

PySpark - Concepts & Understanding

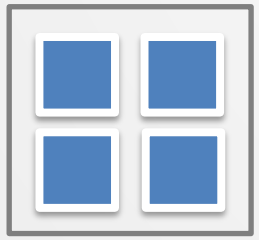
Pandas

Fast, Powerful, Flexible & Easy to Use Open-Source Data Analysis and Manipulation Tool, Built on Top of Python Programming Language. Data Scientists

Additional Information

- Open-Sourced in **2009**
- Fast & Efficient DataFrame Object for Data Manipulation
- Reading & Writing Data ~ In-Memory Data Structures
- Optimized for Performance using Cython & C

Single-Node



Friction

PySpark

Interface for Apache Spark in Python. Allows You to Write Applications using Python APIs. Provides Shell for Interactively Analyzing Your Data in a Distributed Environment

Additional Information

- Open-Sourced in **2016**
- Dependencies ~ Pandas, Numpy, PyArrow
- Scaling Python Application Transparently
- Koalas Launched in **2019**

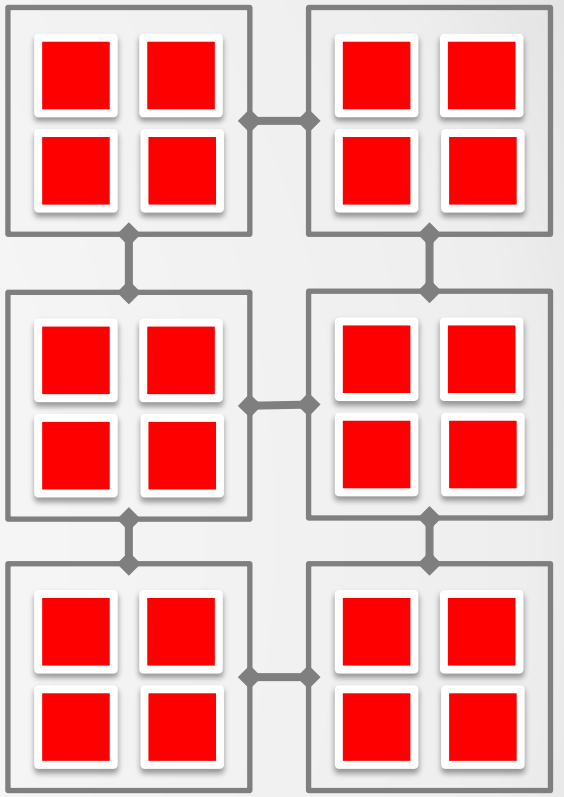
Frictionless

Apache Spark 3.2

Officially Merged into PySpark Engine ~ Project Zen ~ Making Data Science and Data Engineering Easier

Unify Small and Big Data API using PySpark ~ Python

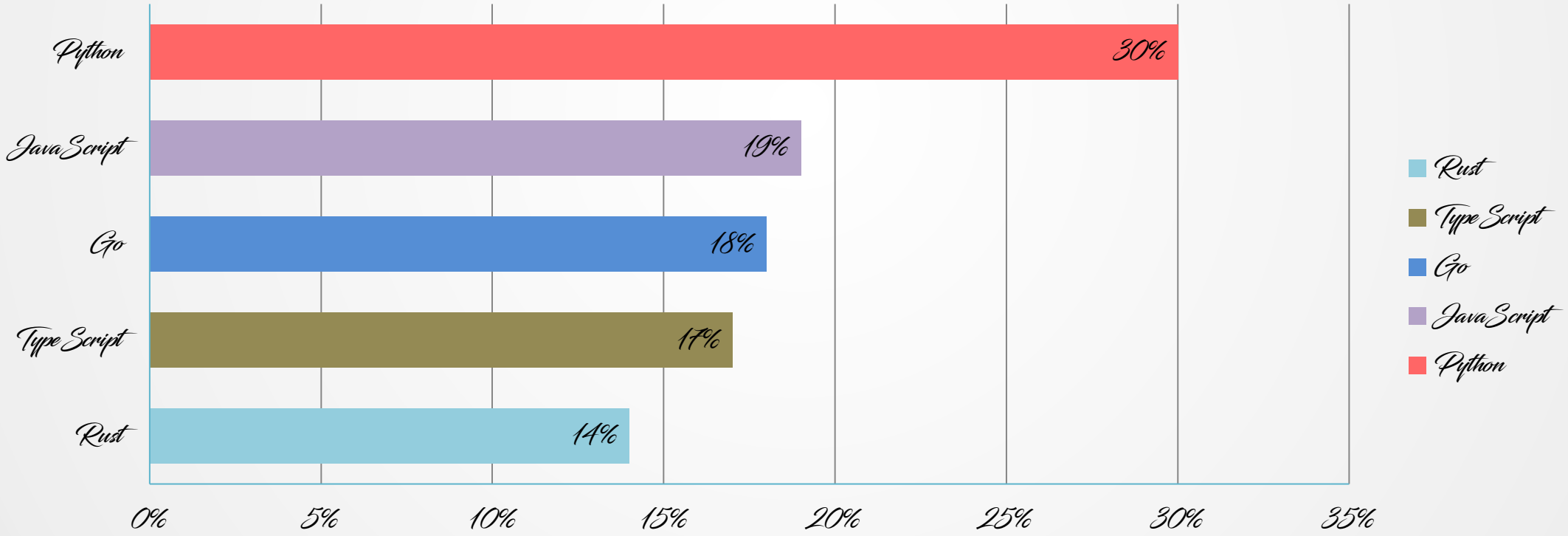
Distributed Computing on Big Data



Stack Overflow Developer Survey 2020 ~ Python

Most Loved, Dreaded, and Wanted Languages

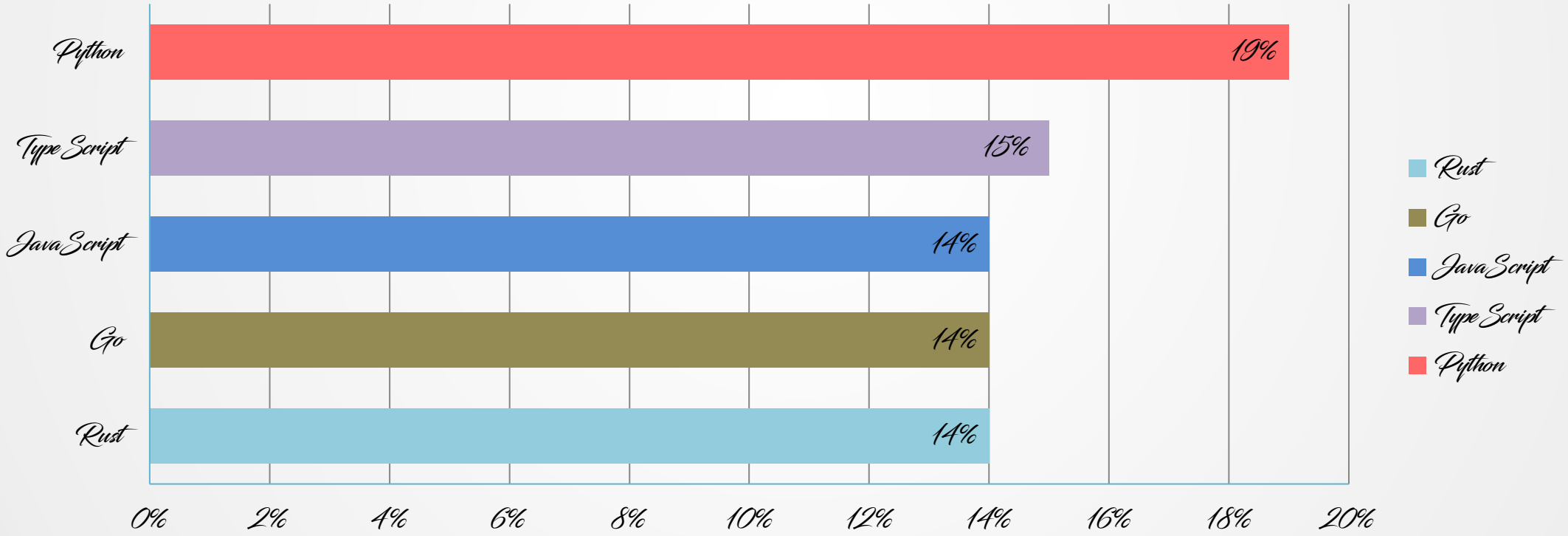
if we look at technologies that developers report that they do not use but want to learn, **python** takes the top spot for the fourth year in a row. we also see some modest gains in the interest in learning rust.



Stack Overflow Developer Survey 2021 ~ Python

Most Loved, Dreaded, and Wanted Languages

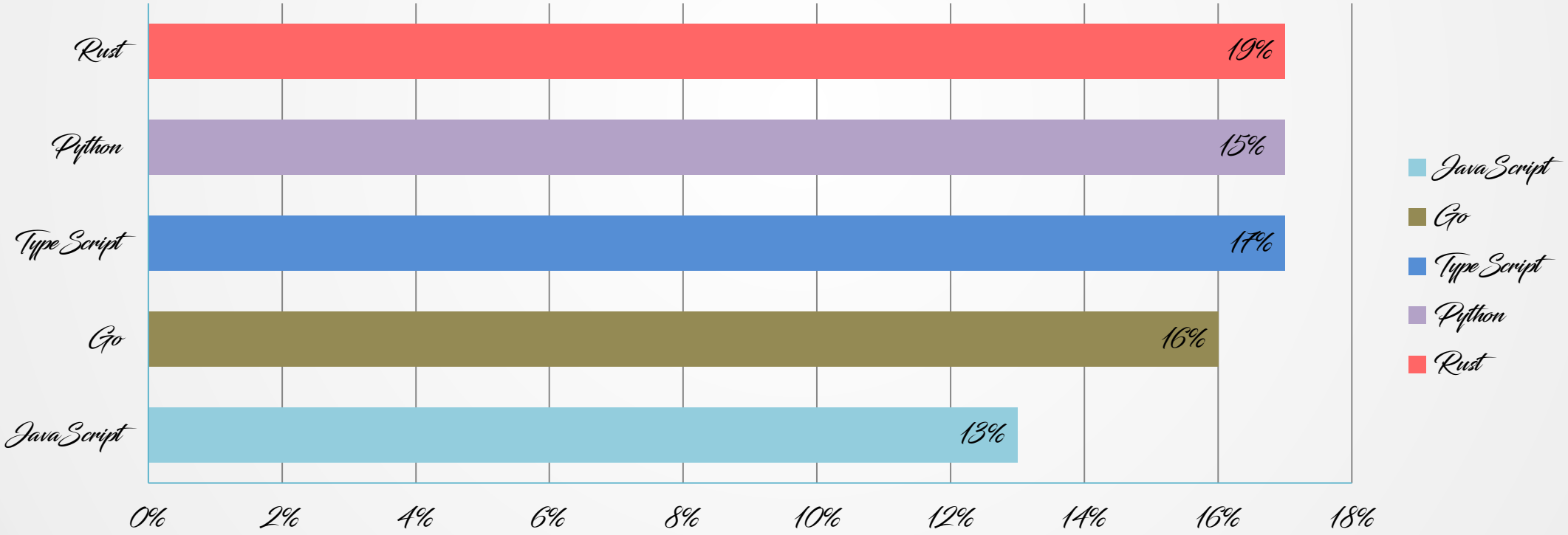
if we look at technologies that developers report that they do not use but want to learn, **python** takes the top spot for the **fifth year in a row**. we also see some modest gains in the interest in learning rust.



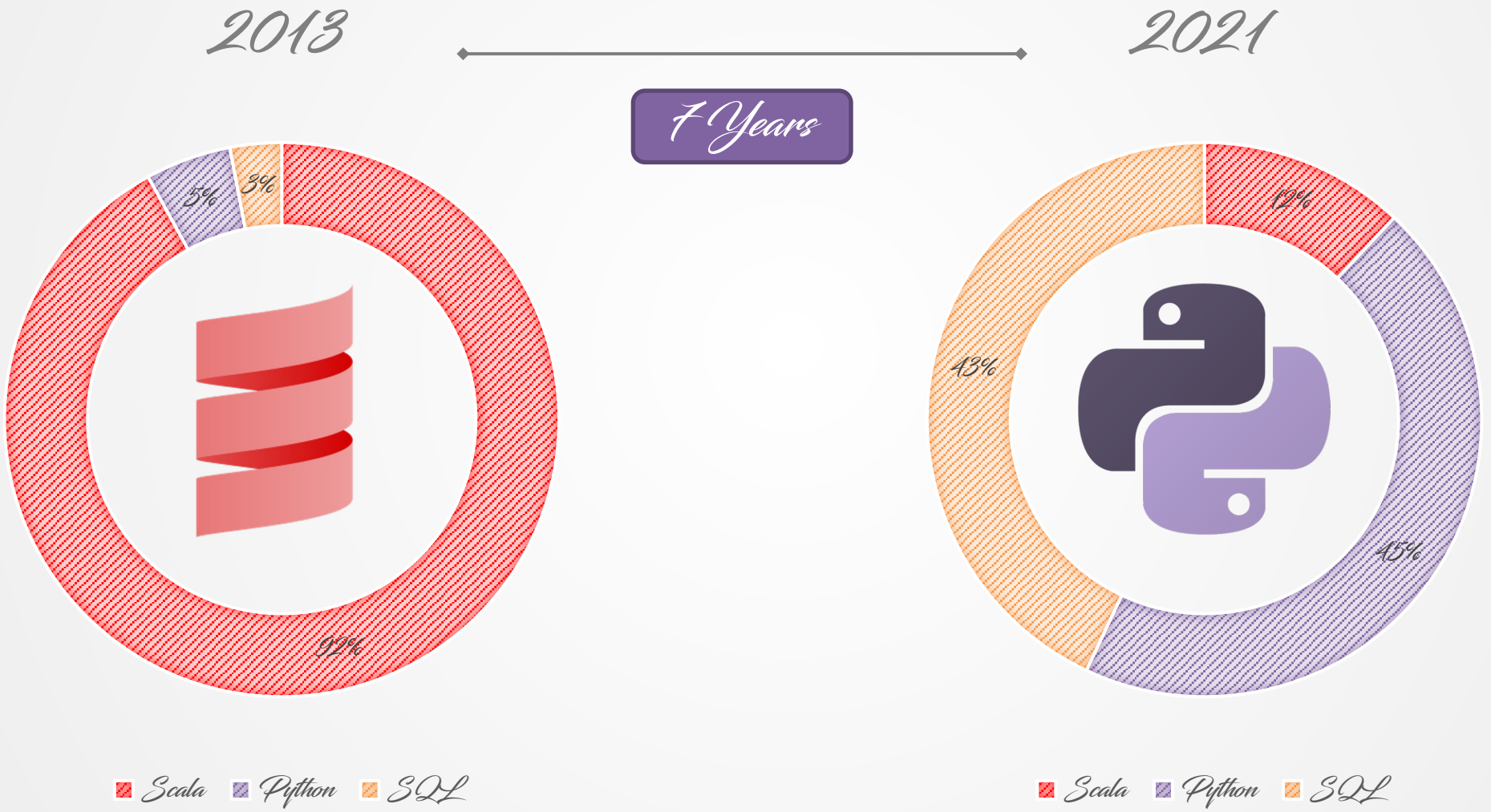
Stack Overflow Developer Survey 2022 ~ Python

Most Loved, Dreaded, and Wanted Languages

if we look at technologies that developers report that they do not use but want to learn, **python** takes the top spot for the fifth year in a row. we also see some modest gains in the interest in learning rust.



PySpark Utilization Report for Apache Spark Workloads



Local Development for Apache Spark

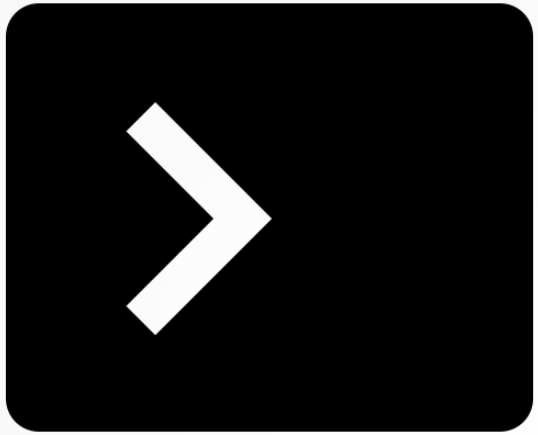
https://spark.apache.org/docs/latest/api/python

PyCharm & PySpark

Interface for Apache Spark in Python. Allows You to Write Applications using Python APIs. Provides Shell for Interactively Analyzing Your Data in a Distributed Environment.

Features Supported

- Core, SQL, DataFrame, Streaming, ML



Spark Shell

Simple Way to Learn Apache Spark's APIs. Powerful Tool to Analyze Data Interactively.

Commands

- *pyspark*
- *sc.version*

```
# Run application locally on 8 cores
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master local[8] \
  /path/to/examples.jar \
  100

# Run on a Spark standalone cluster in client deploy mode
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master spark://207.184.161.138:7077 \
  --executor-memory 20G \
  --total-executor-cores 100 \
  /path/to/examples.jar \
  1000

# Run on a Spark standalone cluster in cluster deploy mode with supervise
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master spark://207.184.161.138:7077 \
  --deploy-mode cluster \
  --supervise \
  --executor-memory 20G \
  --total-executor-cores 100 \
  /path/to/examples.jar \
  1000
```

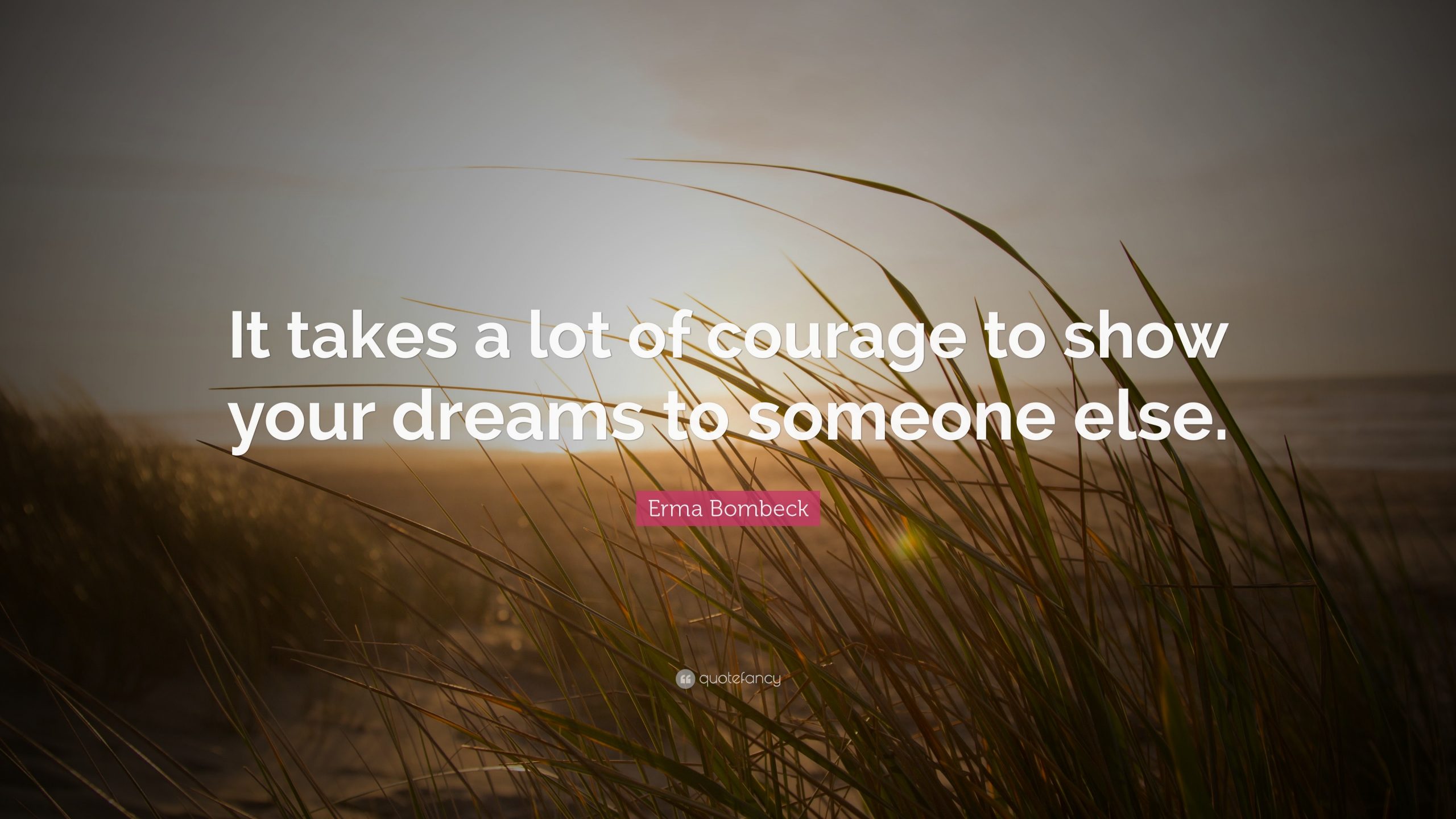
Spark Submit

Used to Launch Applications on a Cluster. Single Script Used to Submit an Apache Spark Program.

Important Notes

- Class ~ Entry Point for Your Application
- Master ~ URL
- Deploy-Mode ~ Client & Cluster
- Conf ~ Configuration Property
- Jars ~ Code Dependencies
- Arguments ~ Passed to Main Method

Introduction and Developing PySpark Program using PyCharm [Local]


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

Apache Spark Deployment Options

Azure HDInsight

Parallel Processing Framework with In-Memory Support to Boost Performance of Big Data Analytics

- HDInsight 4.0
- Ubuntu 16.0.4 LTS
- **Apache Spark 3.1**

Kubernetes

Kubernetes Operator for Apache Spark. Running Spark Applications at Scale with Low and Affordable Cost

- **Apache Spark 3.1.1**

Spark Pools

Microsoft's Implementation of Apache Spark in Azure Synapse Compatible with Azure Storage

- **Apache Spark 3.1**
- Delta Lake Integration
- .NET for Apache Spark

Amazon EMR

Industry-Leading Cloud Big Data Platform for Processing Vast Amount of Data Using Open-Source Tools

- EMR 6.3.0
- Linux AMI
- **Apache Spark 3.1.2**

Databricks

Data + AI Company. Open-Source Apache Spark, Founded in 2013 by Original Creators of Apache Spark

- Apache Spark
- Delta Lake
- MLFlow
- SQL Analytics

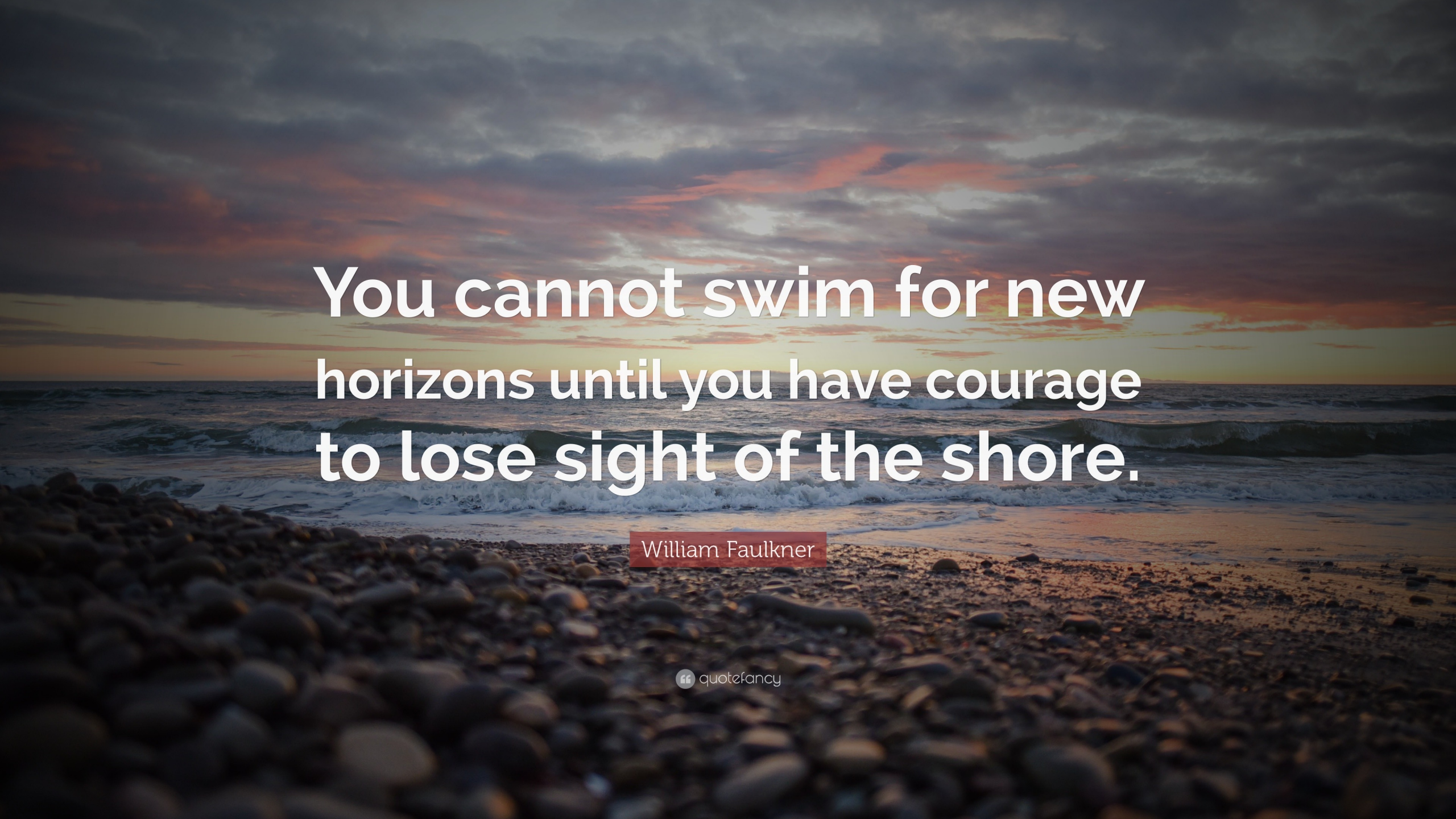
- Runtime 10.4
- **Apache Spark 3.2.1**

Google DataProc

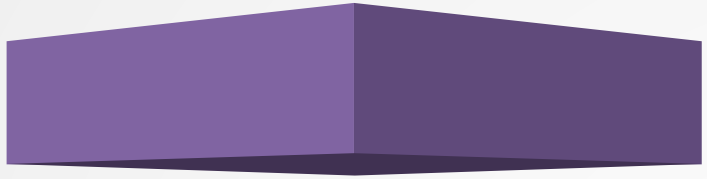
Fully Managed and Highly Scalable Service for Running Apache Spark & +30 OSS

- BORG = Kubernetes
- Debian, Ubuntu & CentOS
- **Apache Spark 3.1.2**

Scaling-Out Spark Applications on Cloud Managed Clusters


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

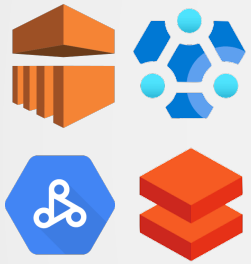
Big Data Eras

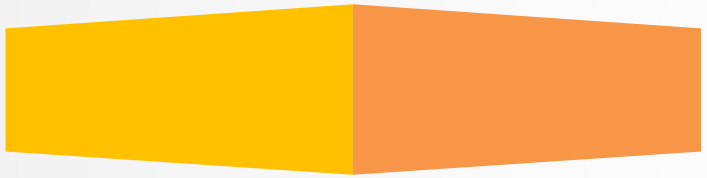


Kubernetes as Cornerstone Solution for Apps & Data ~ [2018]

- Multi-Cloud Strategy First
- Use of Managed Kubernetes Solutions with 99.99%
- Simplify Operations By using Kubernetes as Infrastructure Backbone
- Microservice Driven Approach ~ Containerize Environments
- Cheapest Solution for Application, DataStore & Big Data Solution

Cons ~ Steep Learning Curve, Mindset Change

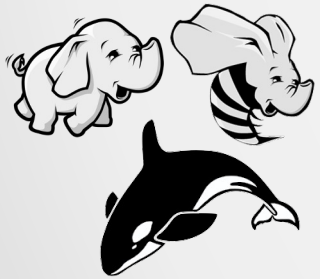


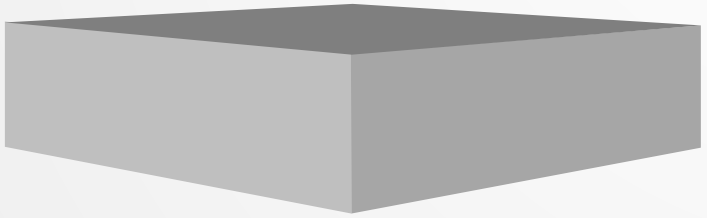


Use of PaaS & SaaS Software's ~ [2014]

- Solution Managed by Cloud Vendor
- Big Data as a Service ~ BDaaS
- Decoupling Computation from Storage
- Pay for Use Only
- Strategy to Reduce Overall Big Data Costs

Cons ~ Promise of Cost Reduction, Myriad of Options





Big Data on Premises using Vendors ~ [2009]

- Cloudera, MapR & HortonWorks
- Based on Open-Source Technologies
- On-Premises Data Centers using Physical Hardware
- Expensive & Dependency Wise
- Large Operations Team ~ Manage Clusters

Cons ~ Expensive, High-Level of Complexity, Burdensome

Kubernetes & Cloud Computing

Open-Source Platform [OSS]

- Kubernetes [K8S]

Microsoft Azure

- AKS – Azure Kubernetes Service

Google Cloud Platform [GCP]

- GKE – Google Kubernetes Engine

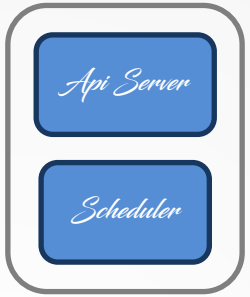
Amazon Web Services [AWS]

- EKS – Elastic Kubernetes Service

Spark-on-Kubernetes ~ Operator

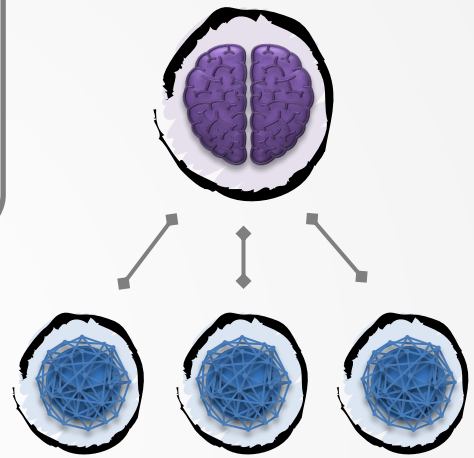
Spark-Submit

Submit an Apache Spark Program
and Launches on Cluster



Spark Driver

Creates Spark Context,
Session, & Executes User Code



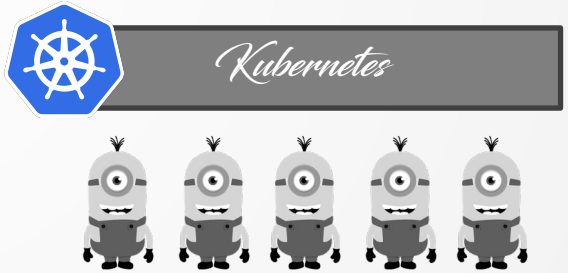
Spark Executor

Executes Individual Tasks &
Return Result Set to Driver

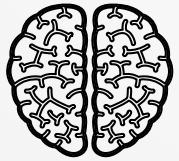
Spark-on-K8s-Operator

Kubernetes Operator ~ Method of Packaging, Deploying and Managing an Application. Extends Functionality to Create, Configure, and Manage Instances of Complex Applications

Spark-on-K8s-Operator ~ Aims to Specifying and Running Apache Spark Applications as Easy and Idiomatic as Running Kubernetes Workloads Instances



spark-pi.yaml



Spark Operator

- Controllers
- Submission Runner
- Pod Monitor
- Mutating Admission

Deployment Mode

Runs in a Kubernetes Pod. Each Worker
Runs Within on Pod Context, use
Kubernetes Master for Cluster Management

Big Data on Kubernetes ~ Spark-on-Kubernetes [Operator]

ONEWAY

SOLUTION