ONEWAY

SOLUTION

One Way Solution

The Data Engineering Pipeline

Data Engineering – [Day 5]

LUAN MORENO

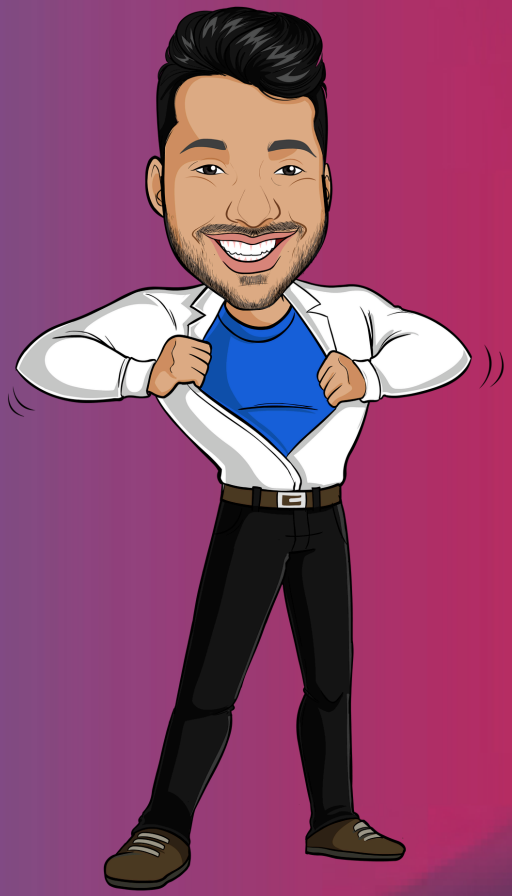
CEO & CDO

Data Engineer & Data Platform MVP

Confluent Certified Developer for Apache Kafka [CCDAK]

Data Warehouse [Dw]

Test Data Engineering – [Day 4]



Data Engineering – Data Warehouse [Dw]

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

Agenda

5

Use-Cases for Apache Spark
The Data Engineering Pipeline
The Data Architecture Landscape
Data Engineering Career

One Way Solution


Success is nothing more than a few simple disciplines, practiced every day.
Jim Rohn
quotefancy

Lambda Architecture – Cloud Agnostic & Simplified

Data Source

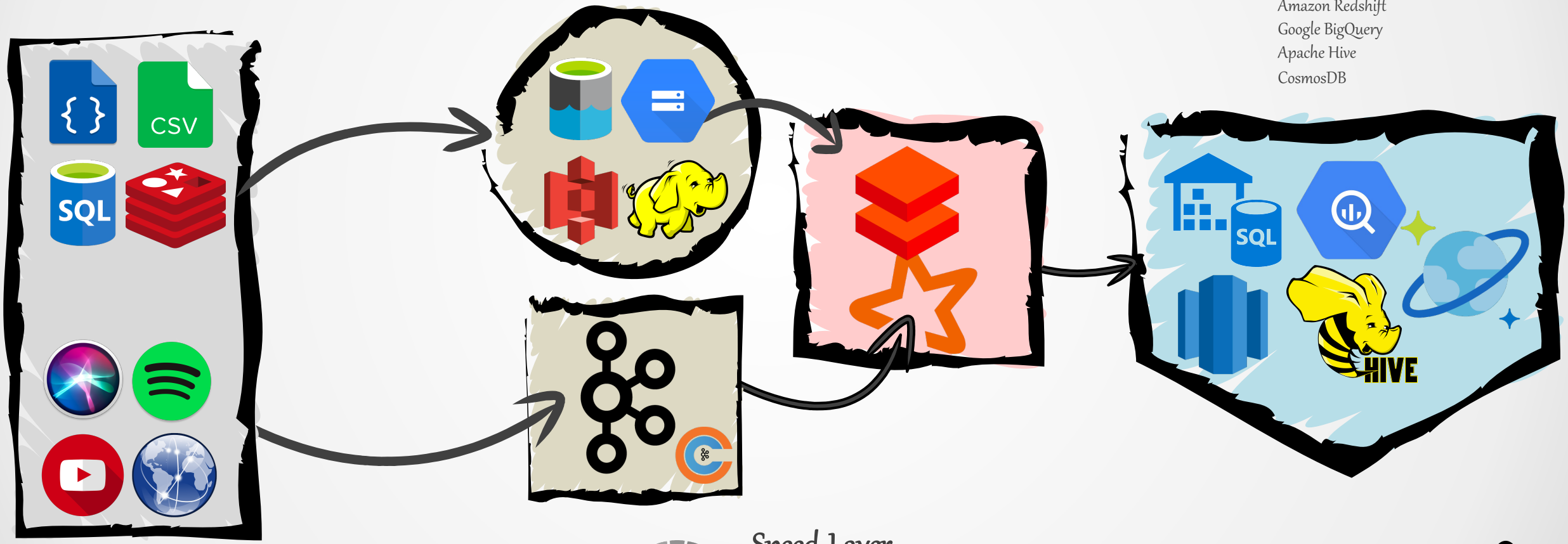
JSON & CSV
SQL Server & Redis
Internet – Siri | Spotify | YouTube

Batch-Layer

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

Serving-Layer

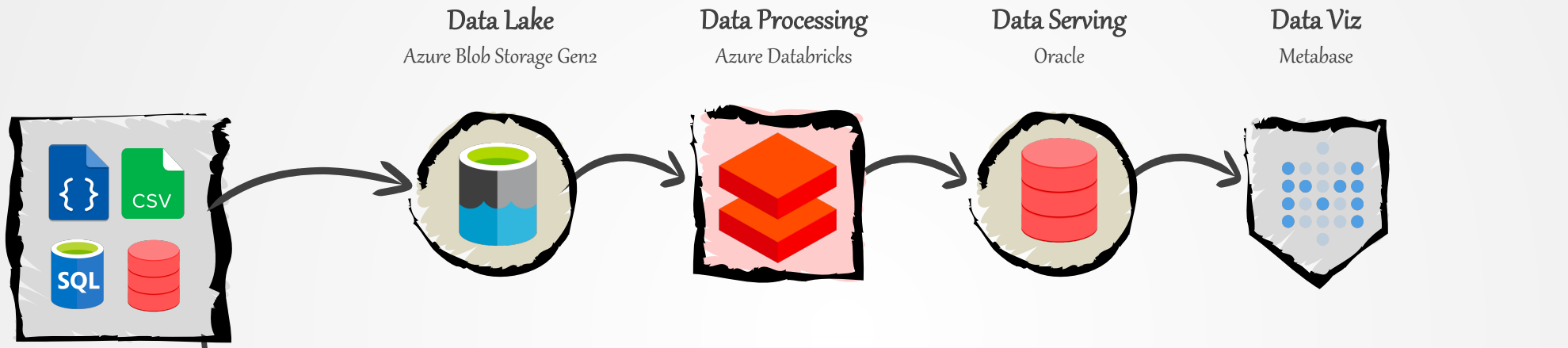
Azure SQL Data Warehouse
Amazon Redshift
Google BigQuery
Apache Hive
CosmosDB



Speed-Layer

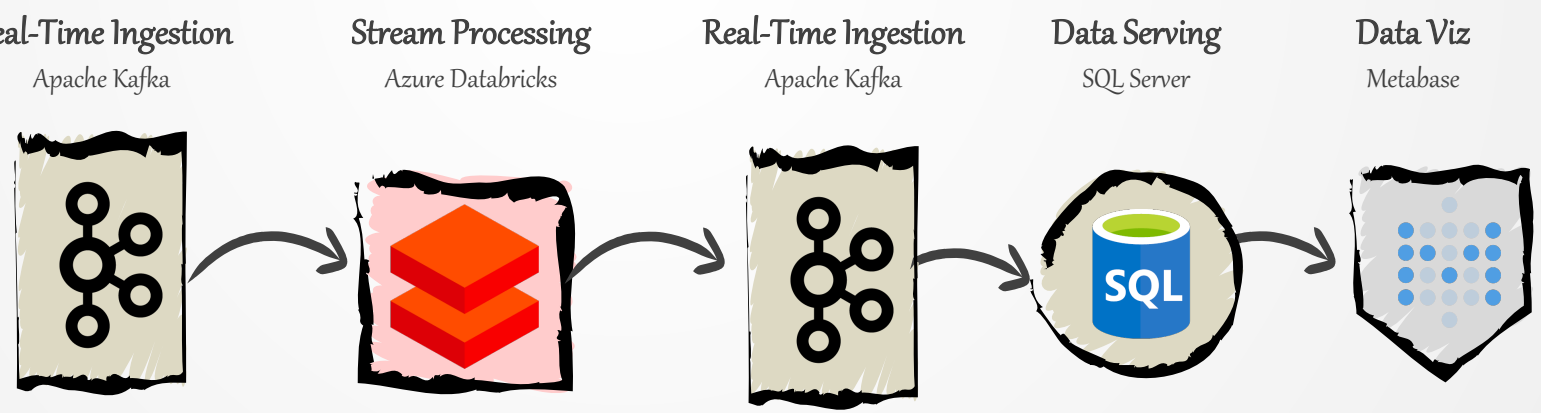
Real-Time Ingestion - Apache Kafka [Confluent]
Stream Processing - Apache Kakfa [Confluent] | Apache Spark [Databricks]

Batch-Layer



Data Sources

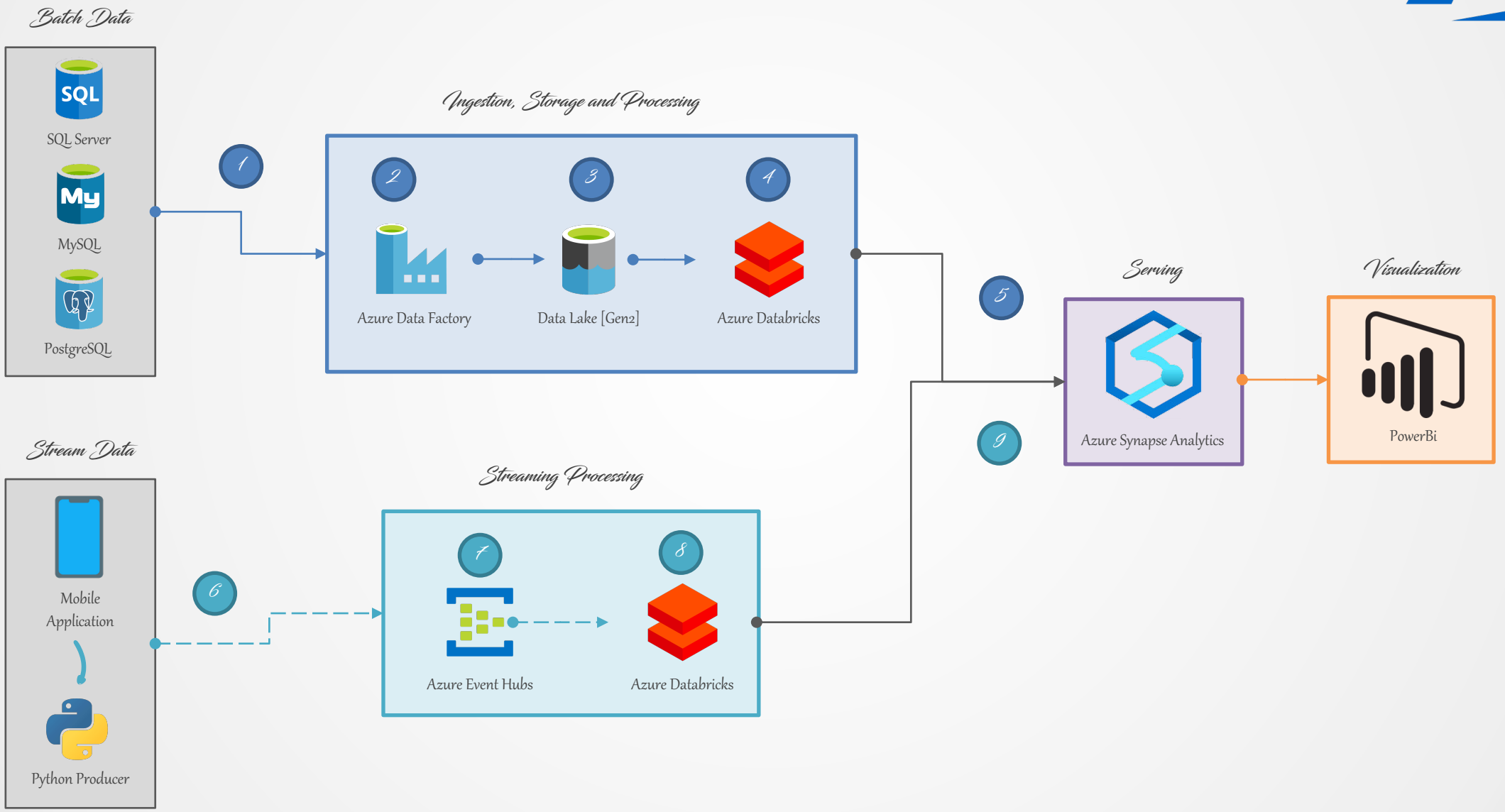
- JSON
- CSV
- RDBMS
- NoSQL



Fast-Layer

Use-Case – Sneak Peak

Lambda Architecture



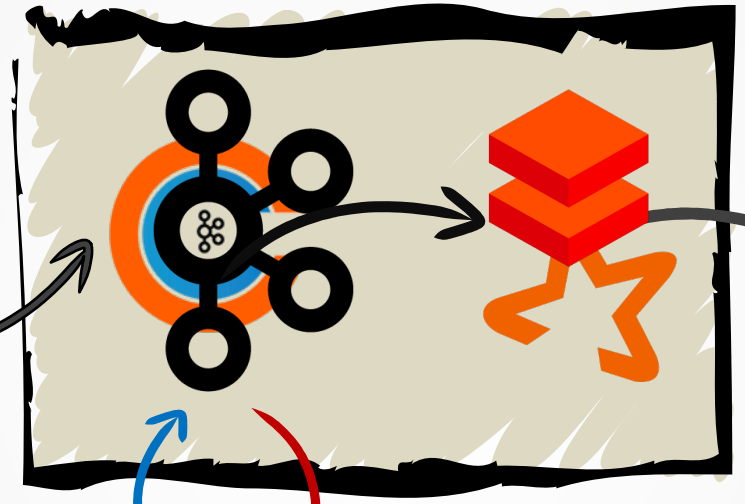
Kappa Architecture – Cloud Agnostic & Simplified

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]

Use-Case – Sneak Peak

Data Sources
RDBMS & NoSQL
Apps & Internet
3rd Party Apps

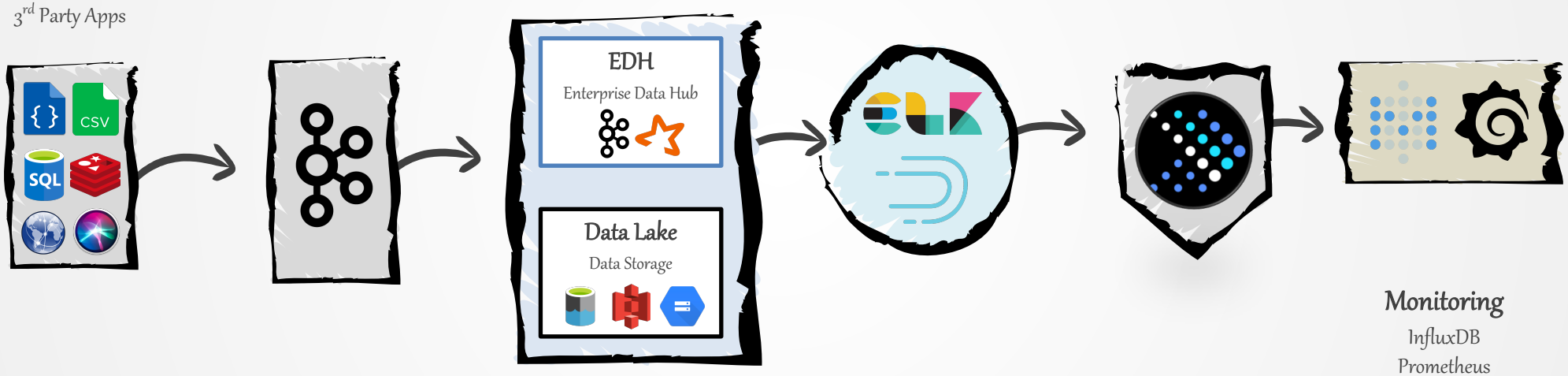
Ingestion
Apache Kafka

CEP
Complex Event Processing

Data Serving
ELK & Druid

Data Integration
Presto

Data Viz
Grafana
Metabase



Monitoring
InfluxDB
Prometheus

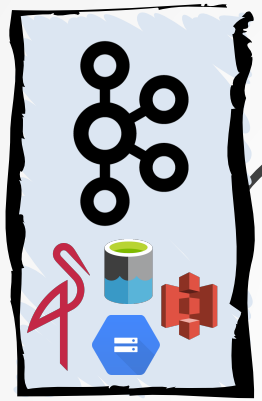
Orion – Big Data as a Service

Kappa Architecture

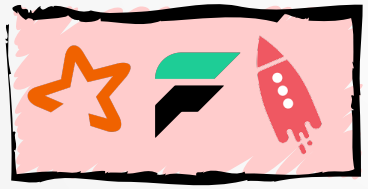
Application Ingestion Layer
Kafka Producer API

Data Store Ingestion Layer
Kafka Connect Source API

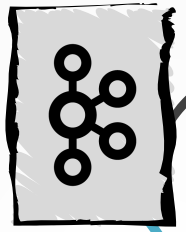
Enterprise Data Hub [EDH]
Apache Kafka with MinIO [Gateways]



Processing
Spark, Faust & KSQLDB



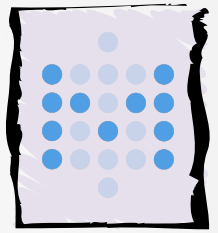
Sink Layer
Kafka Connect Sink API



External Stores
3rd Party Applications

Data Serving
Apache Druid & YugaByteDB

Data Visualization
Metabase



Tracing
Jaeger

Continuous Delivery [CD]
ArgoCD - GitOps

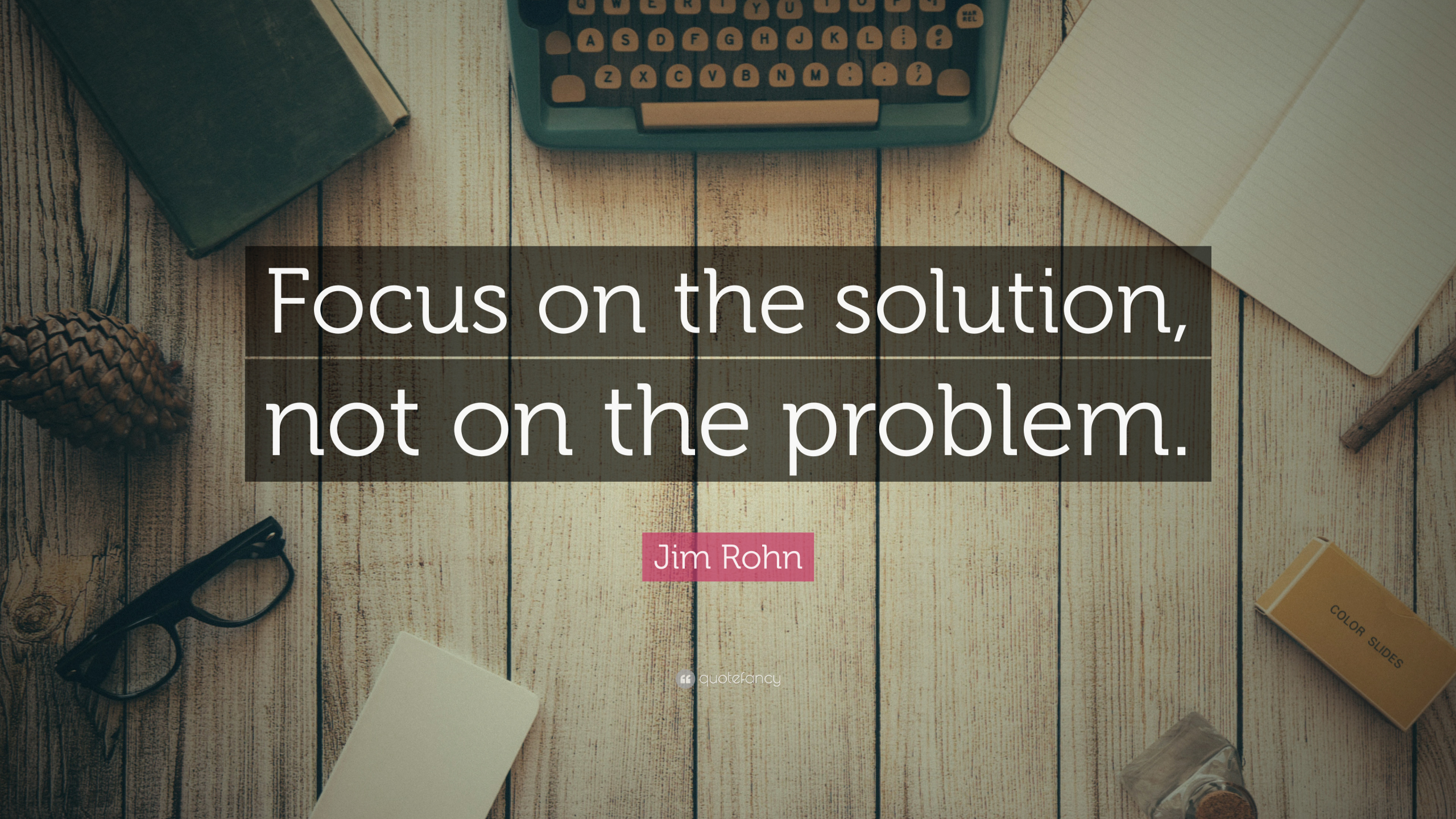
Logging
FileBeat, ElasticSearch & Kibana

Monitoring
Prometheus, Alert Manager & Grafana

AKS
Azure Kubernetes Service

GKE
Google Kubernetes Engine

EKS
Elastic Kubernetes Service


Focus on the solution,
not on the problem.
Jim Rohn
quotefancy

Data Pipeline on Microsoft Azure

Data Ingestion
Azure Data Factory
Azure Event Hubs
HDInsight ~ [Apache Kafka]
Confluent Cloud
Azure Blob Storage

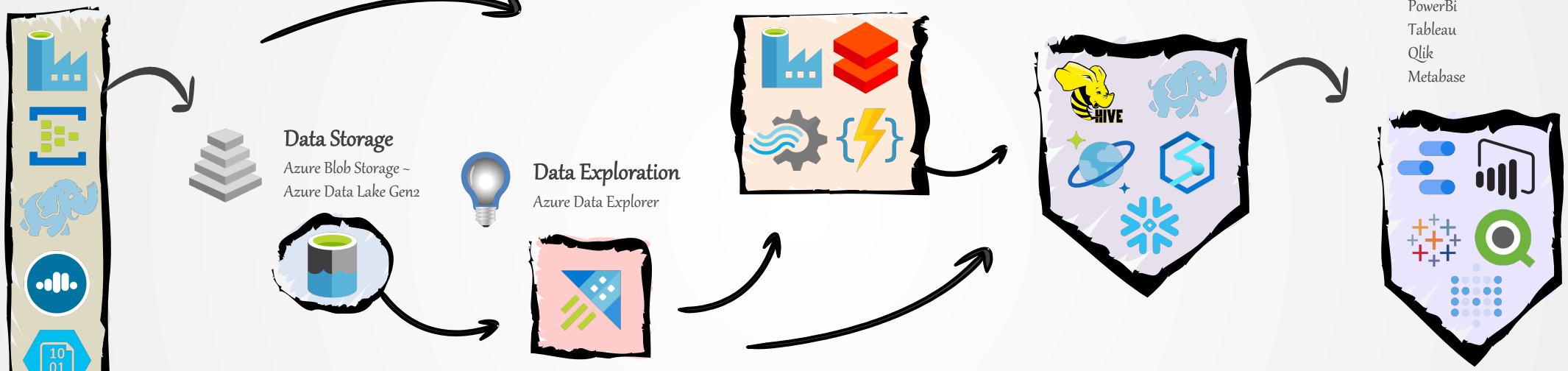
Data Storage
Azure Blob Storage ~
Azure Data Lake Gen2

Data Exploration
Azure Data Explorer

Data Processing
Azure Data Factory ~ [Mapping Data Flows]
Azure Databricks
Azure Stream Analytics
Azure Functions

Data Serving
HDInsight ~ [Apache Hive]
HDInsight ~ [Interactive Query]
Azure CosmosDB
Azure Synapse Analytics
Snowflake

Data Viz
Data Studio
PowerBi
Tableau
Qlik
Metabase



Shared Resources
Shared Among Pipeline

Data Discovery
Azure Data Catalog

RDBMS
Azure SQL DB
Azure DB for MySQL
Azure DB for PostgreSQL

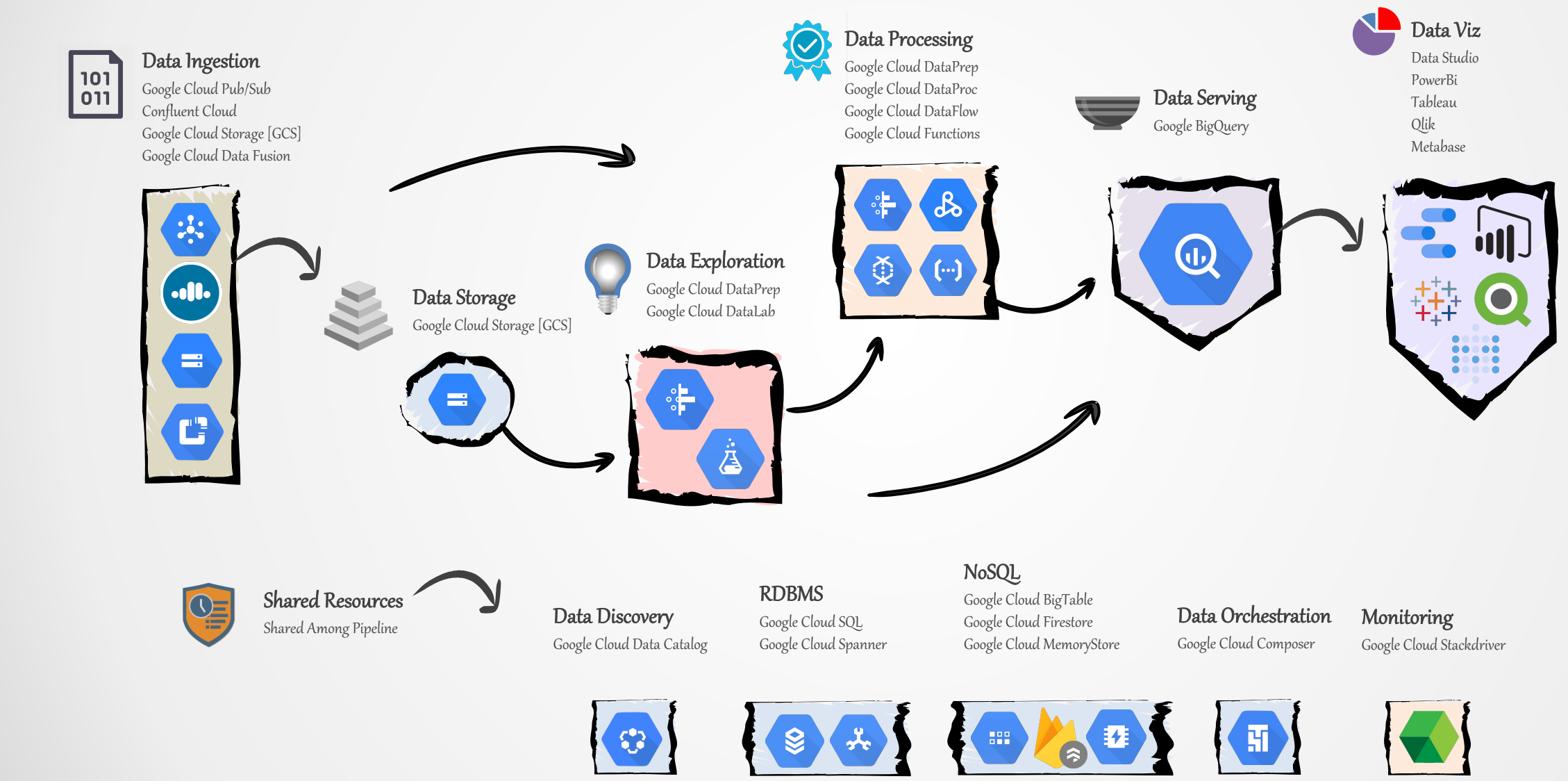
NoSQL
Azure CosmosDB
Azure Cache for Redis

Search
Azure Cognitive Search

Orchestration
Azure Data Factory

Monitoring
Azure Monitor

Data Pipeline on Google Cloud Platform



Data Pipeline on Amazon AWS

Data Ingestion

AWS Data Pipeline
AWS Glue
Kinesis Firehose
Kinesis Data Streams
Amazon MSK
Confluent Cloud
Amazon S3

Data Storage

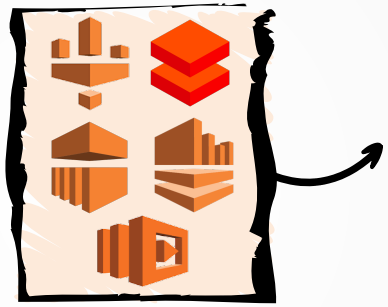
Amazon S3 ~ AWS Lake Formation

Data Exploration

Amazon Athena

Data Processing

AWS Glue ~ DataBrew
Databricks
Amazon EMR
Kinesis Analytics
AWS Lambda



Data Serving

Amazon EMR ~ [Apache Hive]
Amazon EMR ~ [Presto]
Amazon Redshift

Data Viz

Data Studio
PowerBi
Tableau
Qlik
Metabase

Shared Resources

Shared Among Pipeline

Data Discovery

AWS Glue

RDBMS

Amazon RDS

NoSQL

Amazon DynamoDB
Amazon Neptune
Amazon ElastiCache

Search

Amazon CloudSearch

Data Orchestration

AWS Glue

Monitoring

Amazon CloudWatch


Action is the foundational
key to all success.
Pablo Picasso
quotefancy

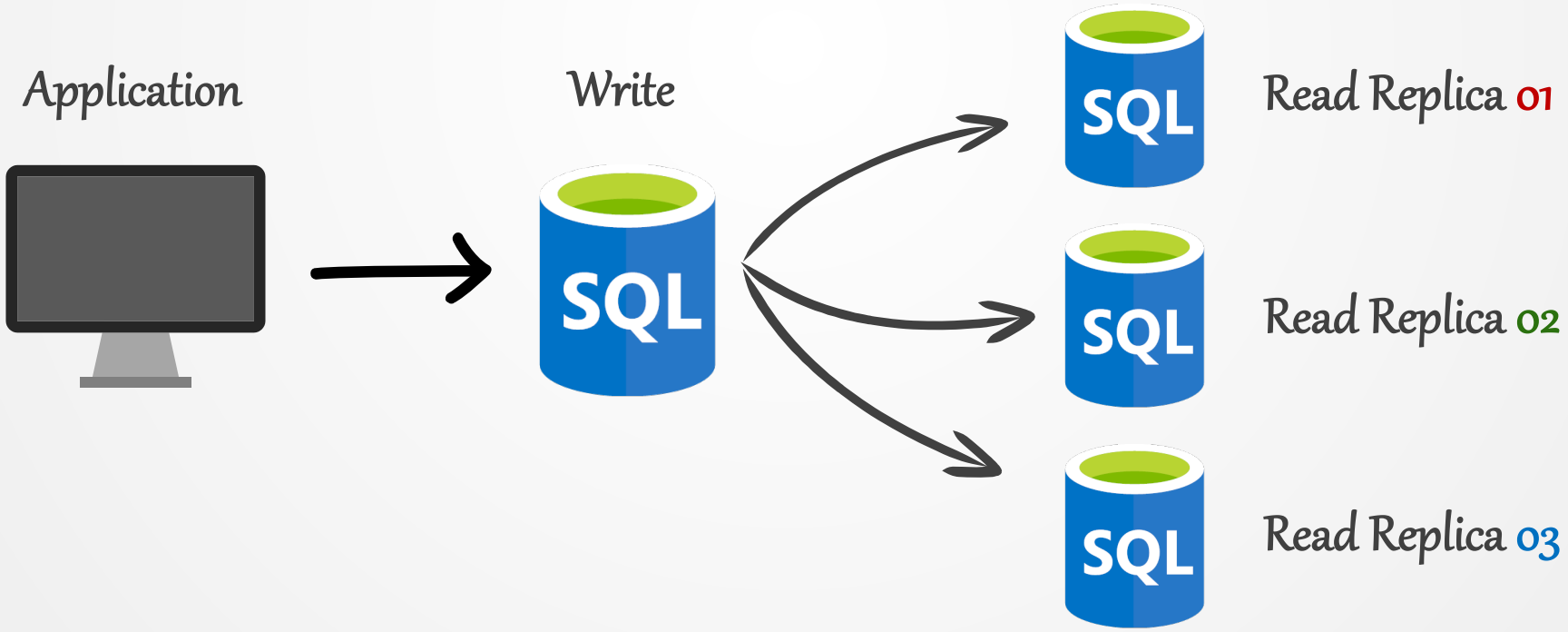
The Data Architecture Landscape

Transactional [OLTP] - 1970

Manage Transaction-Oriented Applications
ACID – Atomicity | Consistency | Isolation | Durability
Vertically Scalable

Know Issues

Concurrency
Write Operations
Scalability



The Data Architecture Landscape

Analytical [OLAP] – 1992 [1.0]

Multi-Dimensional Analytical for BI
Used for Consolidation, Drill-Down, Slicing & Dicing
Complex Analytical & Ad-Hoc Queries

Know Issues

Expensive
Data Silos
Complex

OLTP System

ETL

Extract | Transform | Load

Data Warehouse [Dw]

Bill Inmon [Top-Down]

OLAP Cube

Pre-Aggregated [Viewpoints]

The Data Architecture Landscape

Analytical [OLAP] – 1996 [2.0]

Multi-Dimensional Analytical for BI
Used for Consolidation, Drill-Down, Slicing & Dicing
Complex Analytical & Ad-Hoc Queries

Know Issues

Expensive
Data Silos
Complex

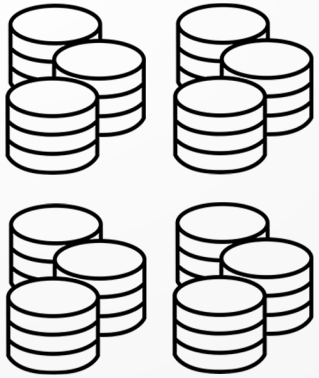
OLTP System

ETL

Extract | Transform | Load

Data Mart [DM]

Departmental Business Decision



Data Warehouse [Dw]

Ralph Kimball [Bottom-Up]

The Data Architecture Landscape

Analytical [OLAP] – 1999 [3.0]

Multi-Dimensional Analytical for BI
Used for Consolidation, Drill-Down, Slicing & Dicing
Complex Analytical & Ad-Hoc Queries

Know Issues

Expensive
Data Silos
Complex

OLTP System

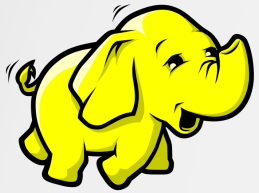
ETL
Extract | Transform | Load

Operational Data Store [ODS]
Departmental Business Decision

ETL
Extract | Transform | Load

Data Warehouse [Dw]
Ralph Kimball [Bottom-Up]

The Data Architecture Landscape

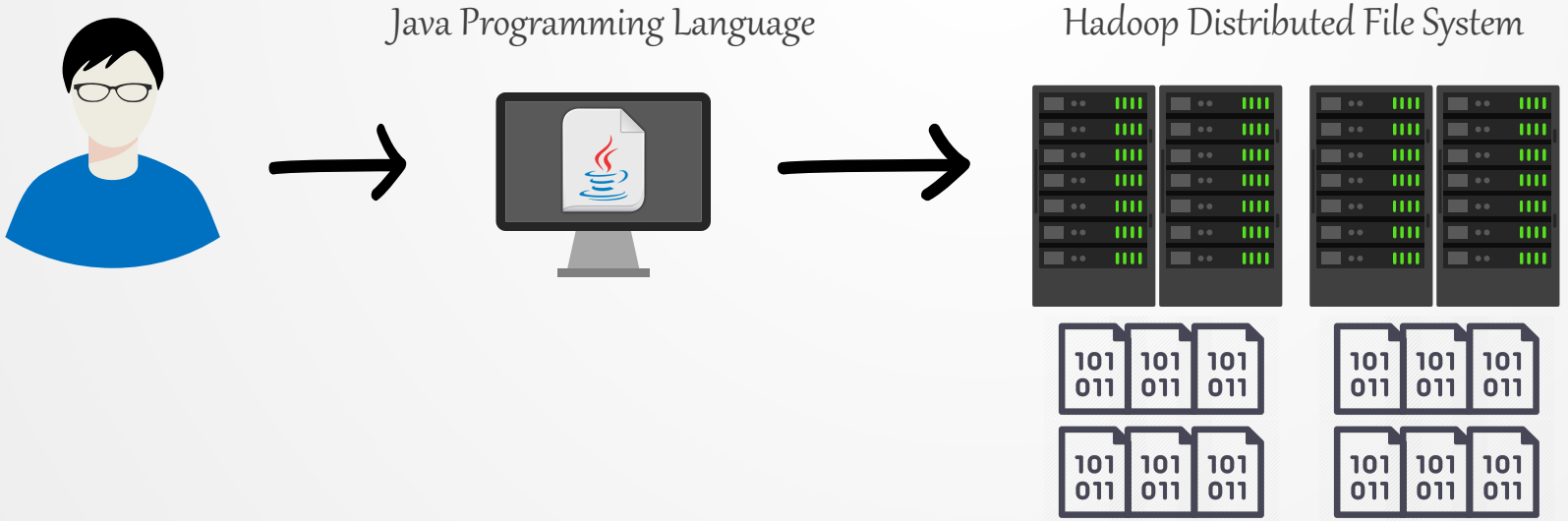


Hadoop – 2006

Cluster of Commodity Hardware
OSS – Open-Source Software
MapReduce + HDFS

Know Issues

Storage Dependent
Java Programming Language
Large Zoo Ecosystem



The Data Architecture Landscape

NoSQL – 2009 [1.0]

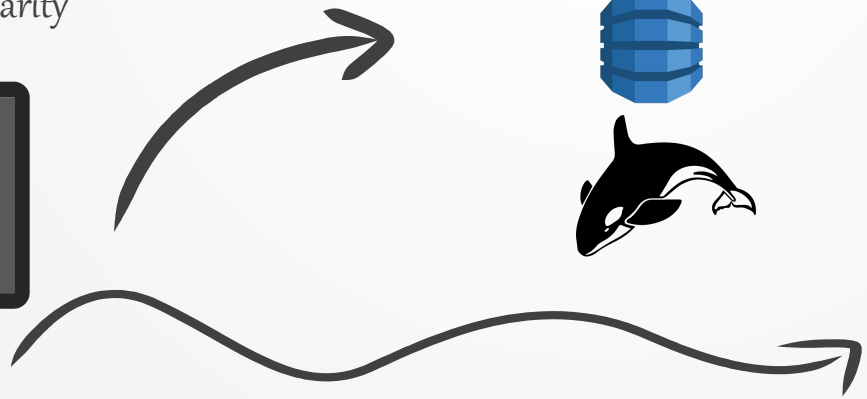
NoSQL - Not Only SQL
Designed for Big Data & Real-Time Web Applications
Types – Key-Value | Wide Column | Document | Graph

Know Issues

Data Consistency
Lack of Standardization
Scalability [Auto-Capabilities]

Monolithic Application

Designed without Modularity



NoSQL

DynamoDB
HBase

SQL

SQL Server

The Data Architecture Landscape

NoSQL + Cloud – 2011 [2.0]

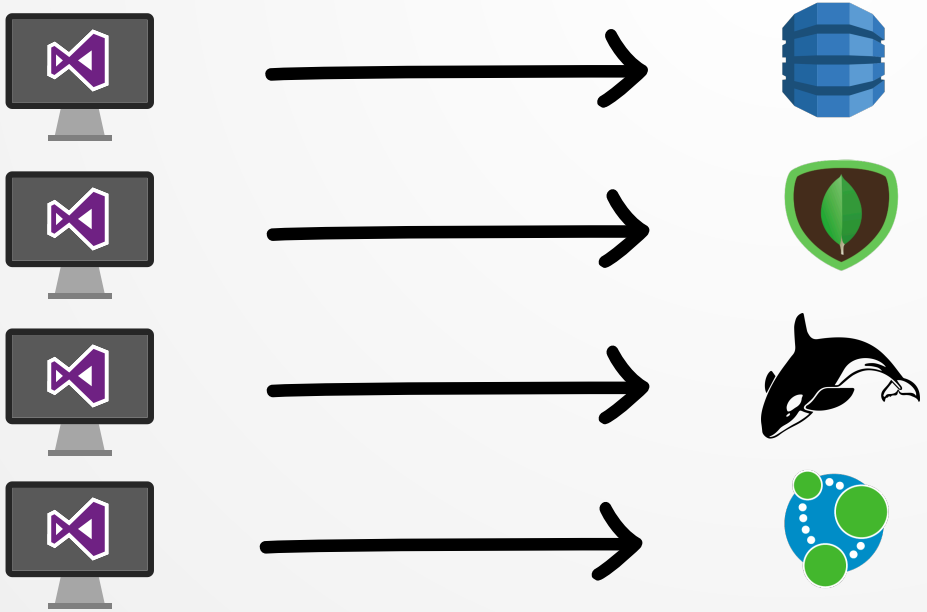
NoSQL - Not Only SQL
Designed for Big Data & Real-Time Web Applications
Types – Key-Value | Wide Column | Document | Graph

Know Issues

Data Consistency
Lack of Standardization
Scalability [Auto-Capabilities]

Micro-Services Application

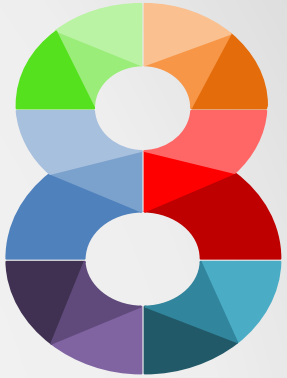
Serve Only One Purpose



NoSQL

DynamoDB | HBase | MongoDB | Neo4J

The Data Architecture Landscape

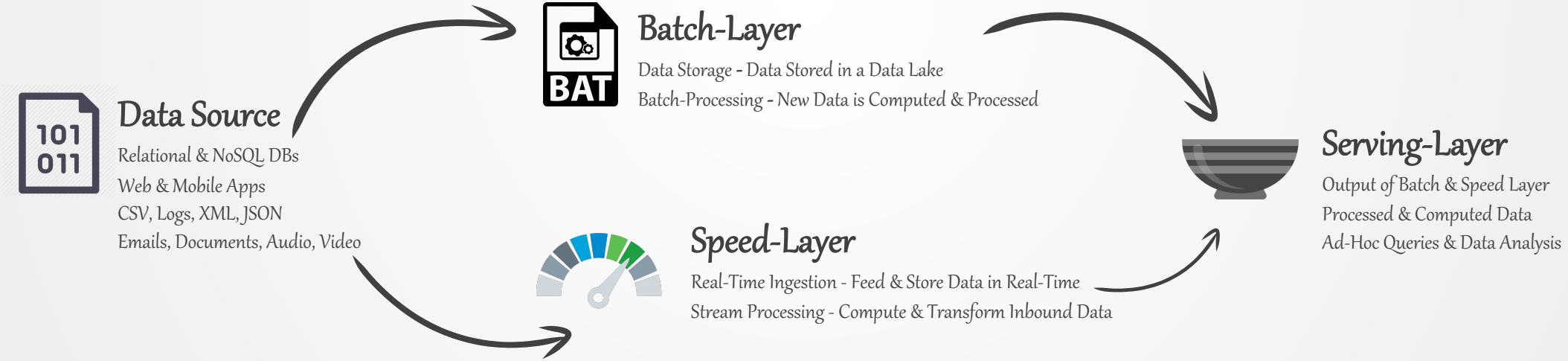


Lambda Architecture – 2013 [1.0]

Data Processing Architecture – Data Pipeline
Designed for Batch & Processing Methods
Big Data & Real-Time Analytics Response

Know Issues

Two Lanes [Batch & Stream]
Maintenance Issues
Variety of Technologies



Data Source

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

Batch-Layer

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

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

The Data Architecture Landscape

Kappa Architecture – 2014 [2.0]

Data Processing Architecture – Data Pipeline
Unify Batch & Stream as an [Event Source]
For Real-Time Applications & Fast Response Times

Know Issues

Leading Edge Technologies
Master Apache Kafka & Apache Spark

Data Source

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

Event-Source

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

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

Long Term Store

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

The Data Architecture Landscape – The State of Art

Kappa Architecture with Steroids – 2020 [3.0]

Data Processing Architecture – Data Pipeline
Unify Batch & Stream as an [Event Source]
State of Art using Delta Lake

Apache Kafka

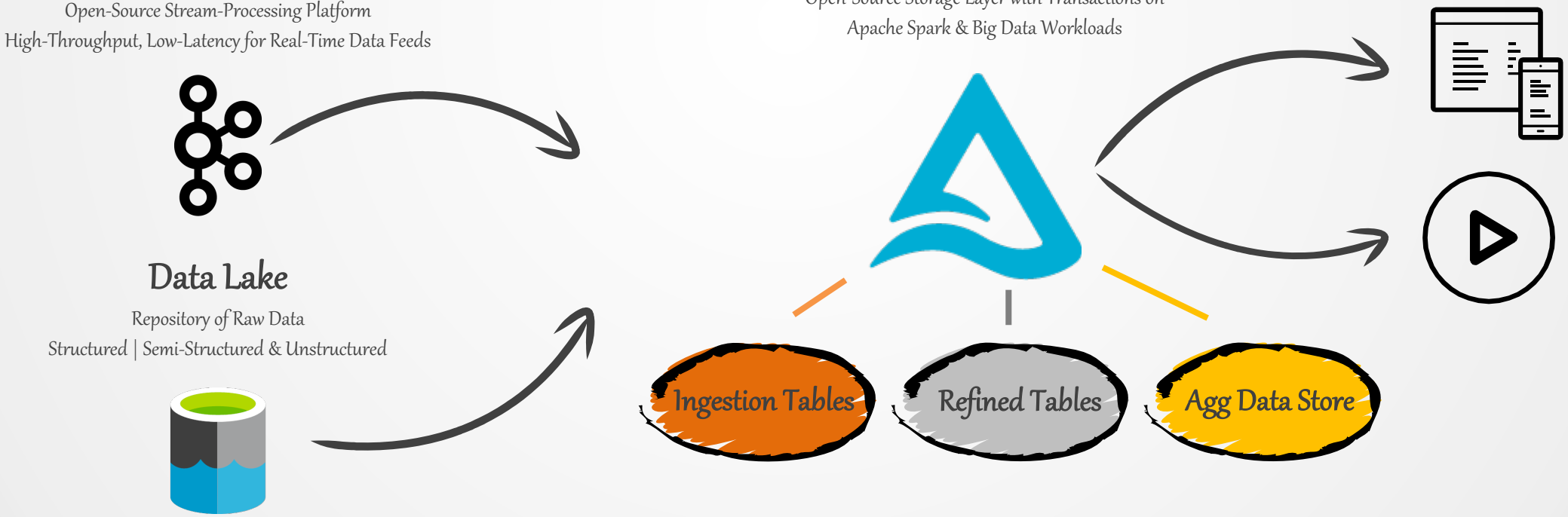
Open-Source Stream-Processing Platform
High-Throughput, Low-Latency for Real-Time Data Feeds

Data Lake

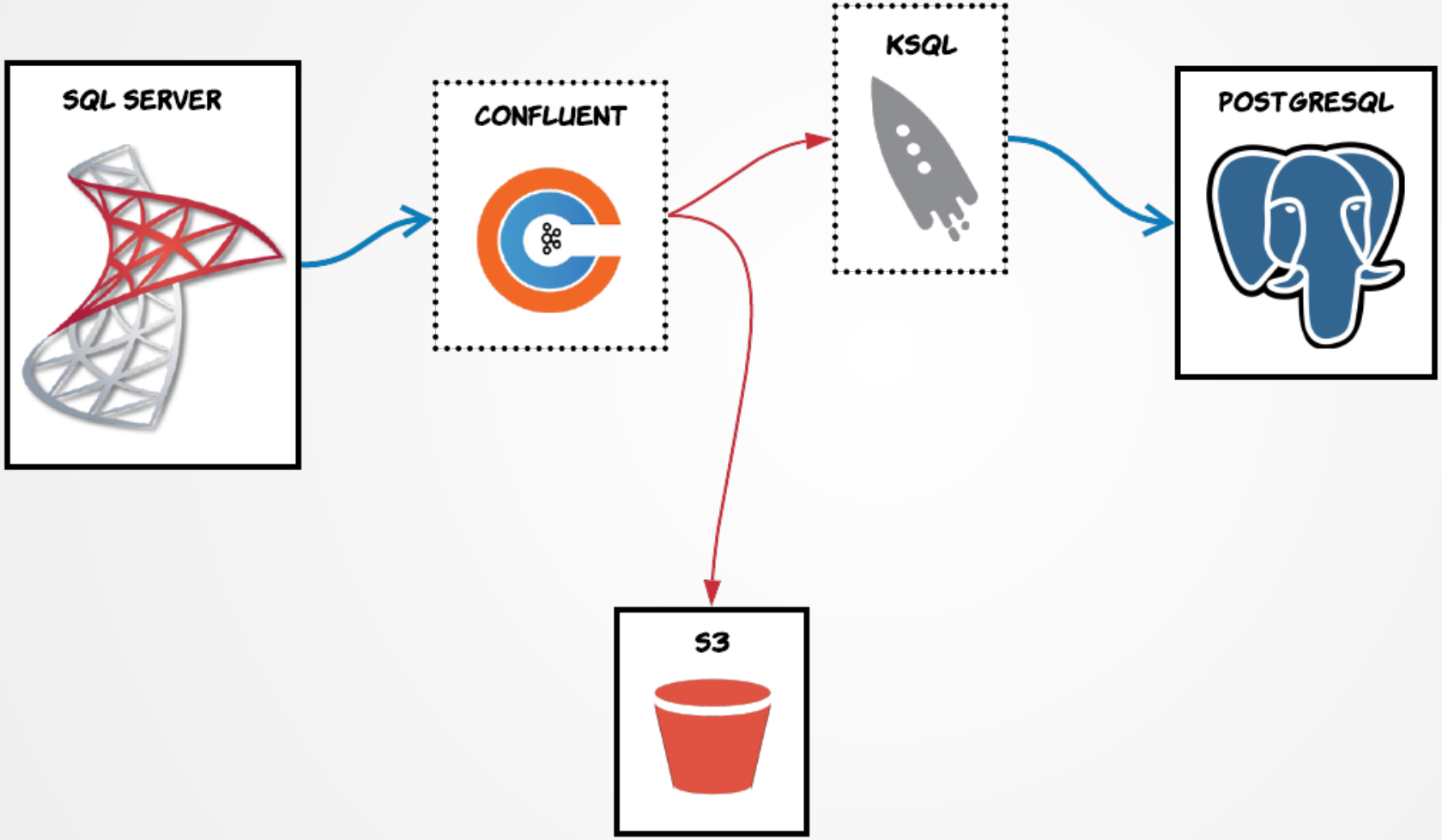
Repository of Raw Data
Structured | Semi-Structured & Unstructured

Data Lakehouse [Delta Lake]

Open-Source Storage Layer with Transactions on
Apache Spark & Big Data Workloads



The Data Architecture Landscape — The State of Art



The Data Architecture Landscape — The State of Art

SQL SERVER

APACHE KAFKA

DATABRICKS

S3

APP

BRONZE

SILVER

GOLD

RAW

DW

DATASET

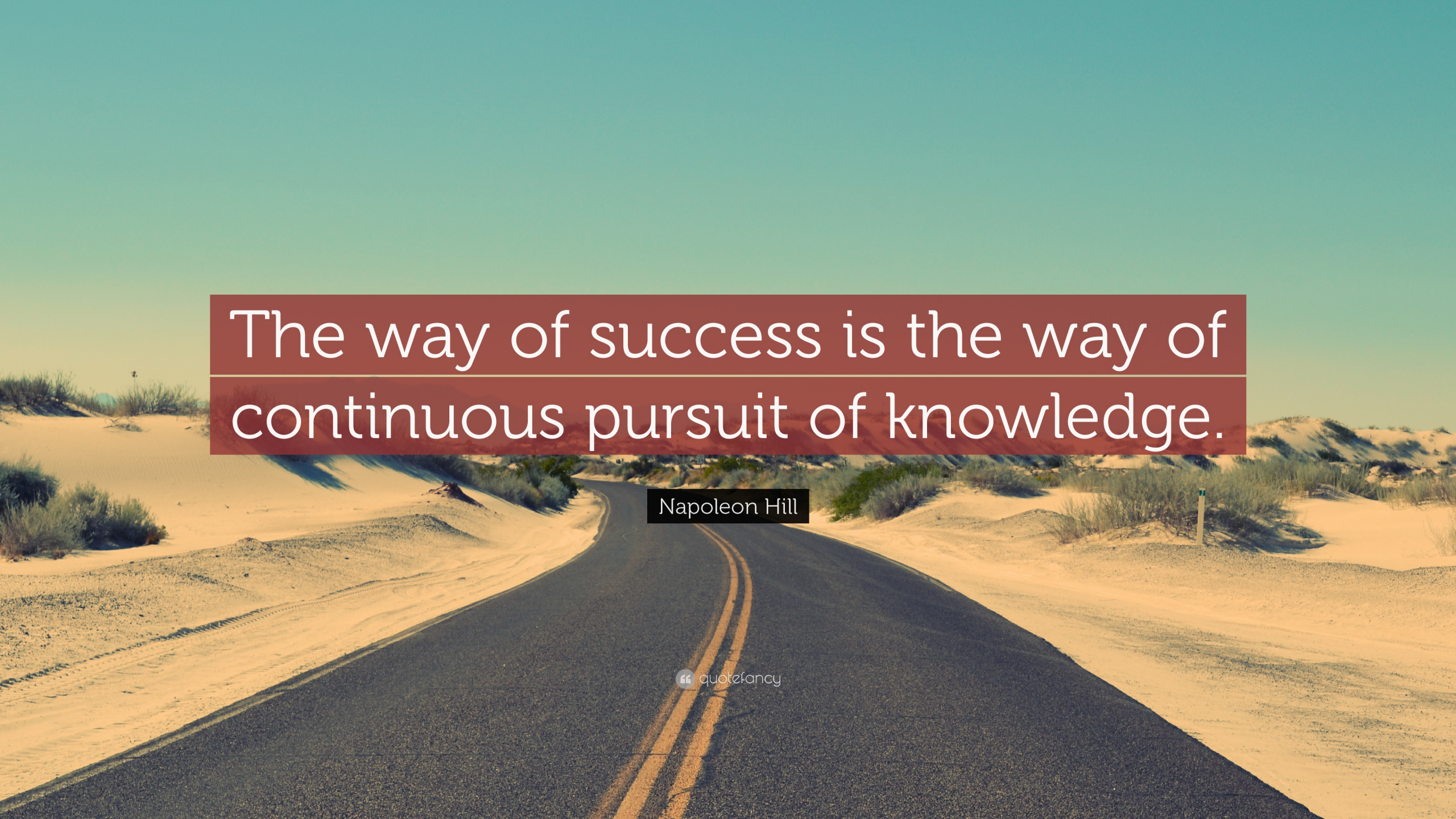
mlflow

K

dmlc XGBoost

H2O.ai

HOROVOD


The way of success is the way of continuous pursuit of knowledge.
Napoleon Hill
quotefancy

“War is 90% information.”

- Napoleon Bonaparte

“A scientist can discover a new star, but he cannot make one. He would have to ask an engineer to do it for him.”

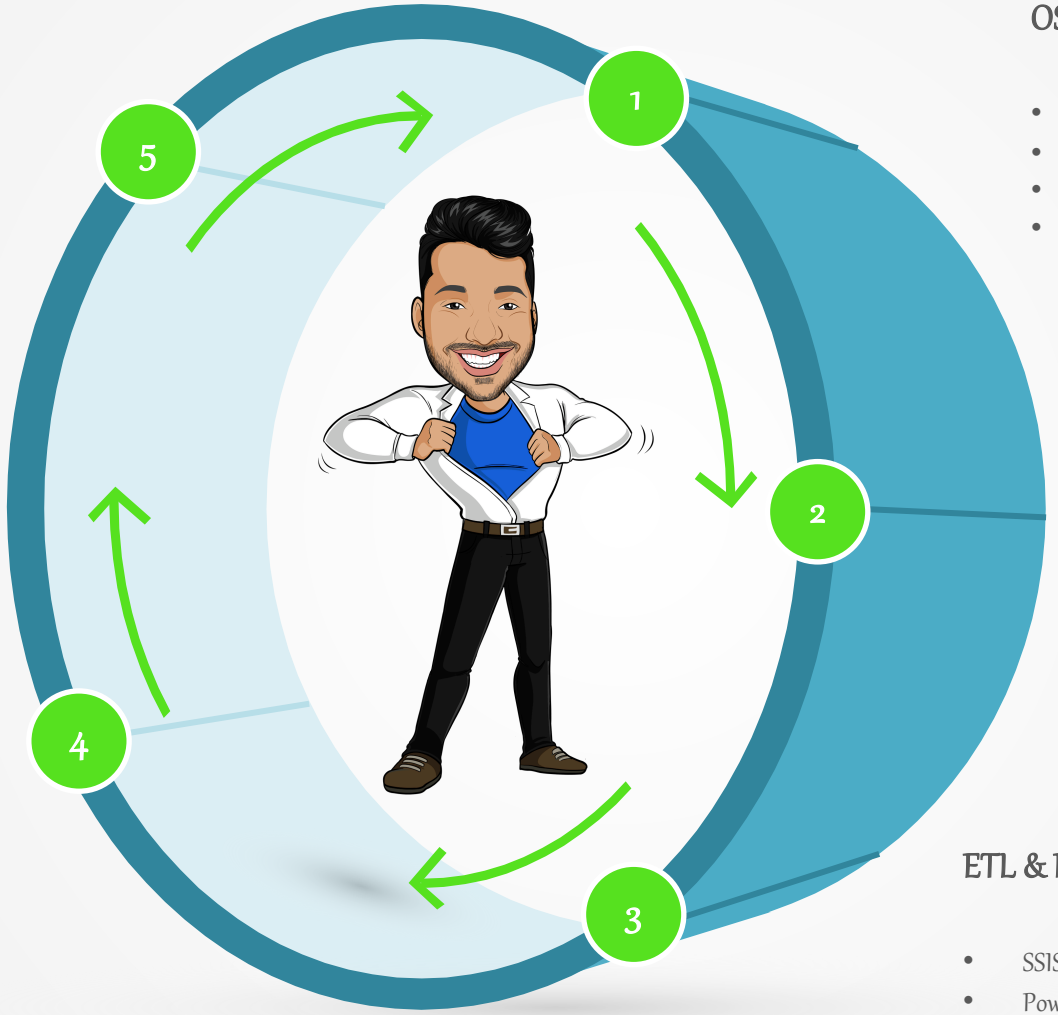
- Gordon Lindsay Glegg

“In God we trust. All others must bring data.”

- W. Edwards Deming

Data Engineer Technical Skills

Data Engineer Career - Part 1



OS & Programming Language

- Linux
- SQL
- Python
- Scala

Data Pipelines & Cloud Computing

- Lambda & Kappa
- Google GCP
- Amazon AWS
- Microsoft Azure

DBMS & NoSQL

- SQL Server
- Oracle
- PostgreSQL
- MySQL
- MongoDB
- Cassandra
- Redis Cache

Distributed Systems & Big Data Frameworks

- Apache Hadoop [HDFS]
- Apache Spark
- Apache Kafka

ETL & DW

- SSIS & ODI
- PowerCenter
- Talend
- Pentaho
- Oracle Exadata
- Sybase IQ

Data Engineer Business Skills

Data Engineer Career - Part 2

01

Creative Problem-Solving

approaching data organization challenges with a clear eye on what is important; employing the right approach/methods to make the maximum use of time and human resources.

02

Effective Collaboration

carefully listening to management, data scientists and data architects to establish their needs.

03

Intellectual Curiosity

exploring new territories and finding creative and unusual ways to solve data management problems.

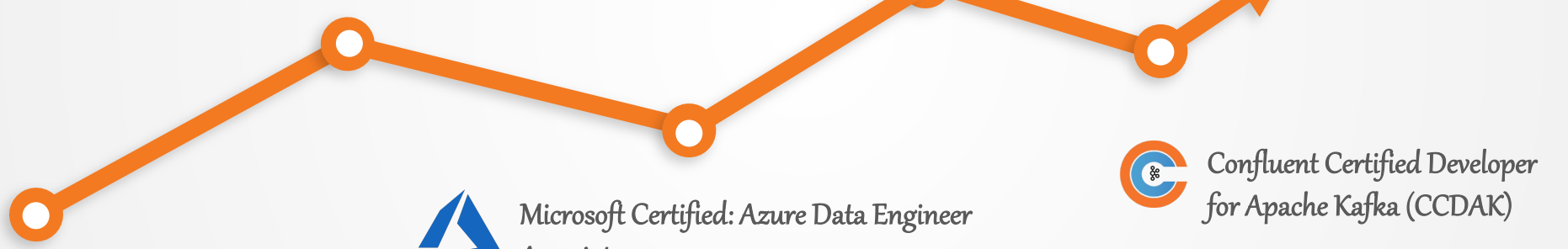
04

Industry Knowledge

understanding the way your chosen industry functions and how data can be collected, analyzed and utilized; maintaining flexibility in the face of big data developments.

Data Engineer Certifications

Data Engineer Career - Part 3



Amazon Web Services (AWS) Certified Big Data – Specialty

The AWS Certified Big Data – Specialty certification is intended for individuals who perform complex Big Data analyses with at least two years of experience using AWS technology.

Databricks Certified Associate Developer for Apache Spark 3.0

Validates your knowledge of the core components of the DataFrames API and confirms that you have a rudimentary understanding of the Spark Architecture.

Google Professional Data Engineer

A Professional Data Engineer enables data-driven decision making by collecting, transforming, and publishing data.

Microsoft Certified: Azure Data Engineer Associate

Azure Data Engineers design and implement the management, monitoring, security, and privacy of data using the full stack of Azure data services to satisfy business needs.

Confluent Certified Developer for Apache Kafka (CCDAK)

This examination is based upon the most critical job activities that a Confluent Developer performs.

Data Engineer Study

Data Engineer Career - Part 4

Spark 3.0 & Big Data Essentials with Scala

now with spark 3.0: learn practical big data with spark dataframes, datasets, rdd and spark sql

Functional Programming in Scala Specialization

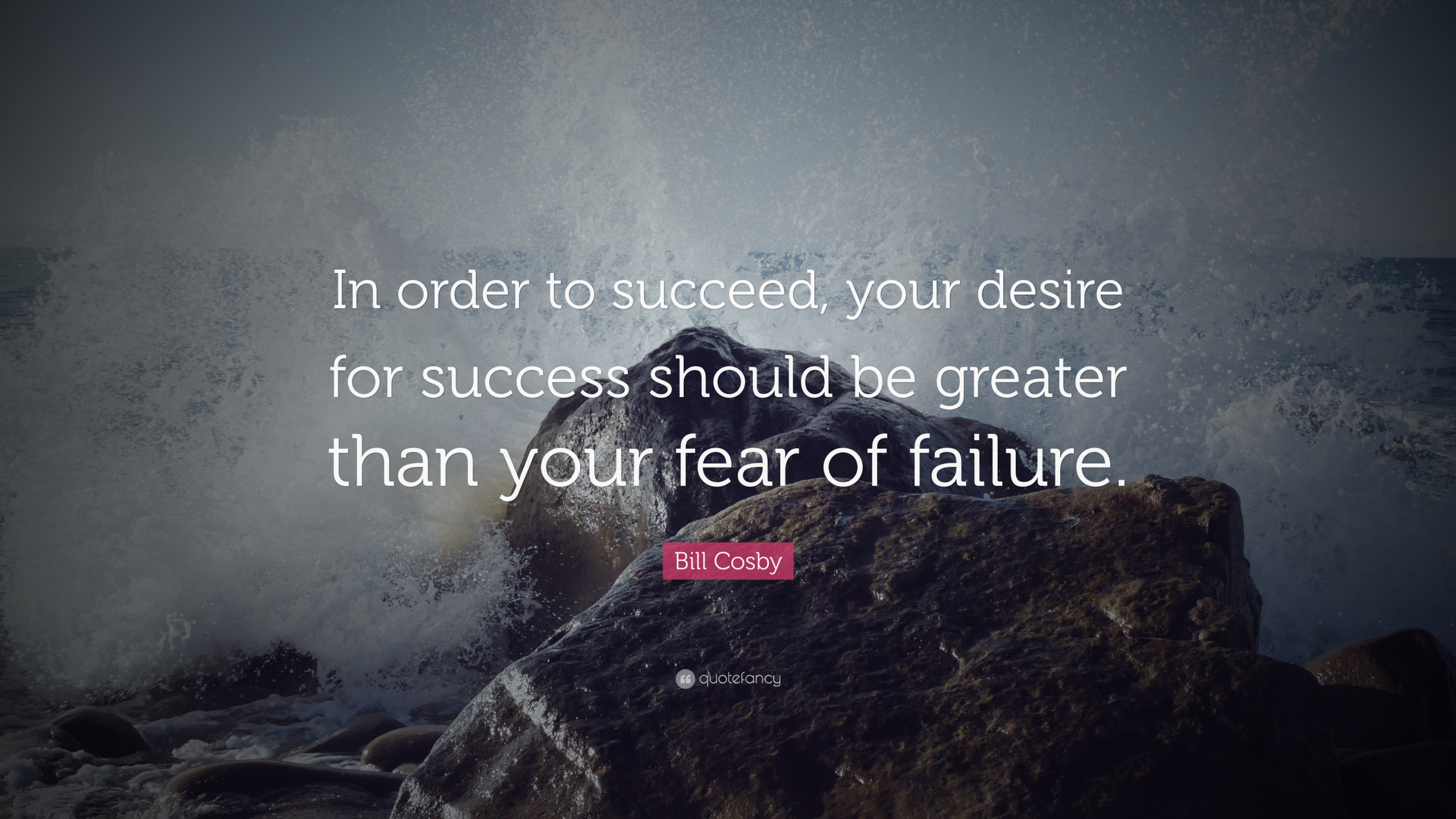
functional programming is becoming increasingly widespread in industry. this trend is driven by the adoption of scala as the main programming language for many applications.

Spark Streaming 3.0 with Scala

stream real time big data with spark 3.0 and integrate with kafka, jdbc, cassandra and other data sources

Data Engineer with Python

an increasing number of businesses are collecting huge amounts of data. Often, they'll have data scientists or data analysts analyze that data


In order to succeed, your desire for success should be greater than your fear of failure.
Bill Cosby
quotefancy

Luan Moreno M. Maciel

YouTube
luanmorenommaciel

LinkedIN
Luan Moreno Medeiros Maciel

Facebook
Luan Moreno Medeiros Maciel

Instagram
engenhariadedados

Meetup
BSB-AI-Big-Data-Analytics

Thank You

One Way Solution

ONEWAY

SOLUTION