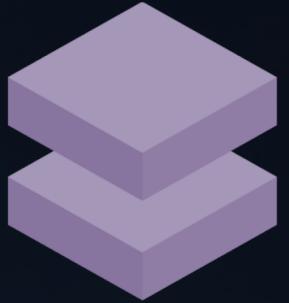




ONEWAY
SOLUTION



One Way Solution

The Data Engineering Pipeline

Data Engineering – [Day 5]

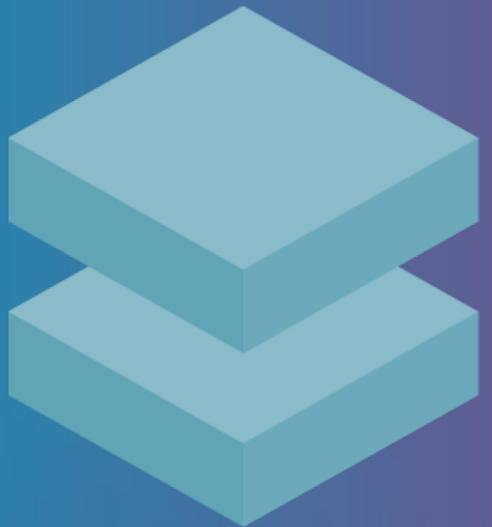


LUAN MORENO
CEO & CDO
Data Platform MVP
Data Architect & Engineer

Data Warehouse [Dw]



Test Data Engineering – [Day 4]



Data Engineering – Data Warehouse [Dw]



4

Traditional Data Warehouse [TDW]

Modern Data Warehouse

Azure SQL DB

Azure Synapse Analytics

Amazon Redshift

Google BigQuery

Apache Hive

Apache Druid

Delta Lake

Use-Cases





Agenda

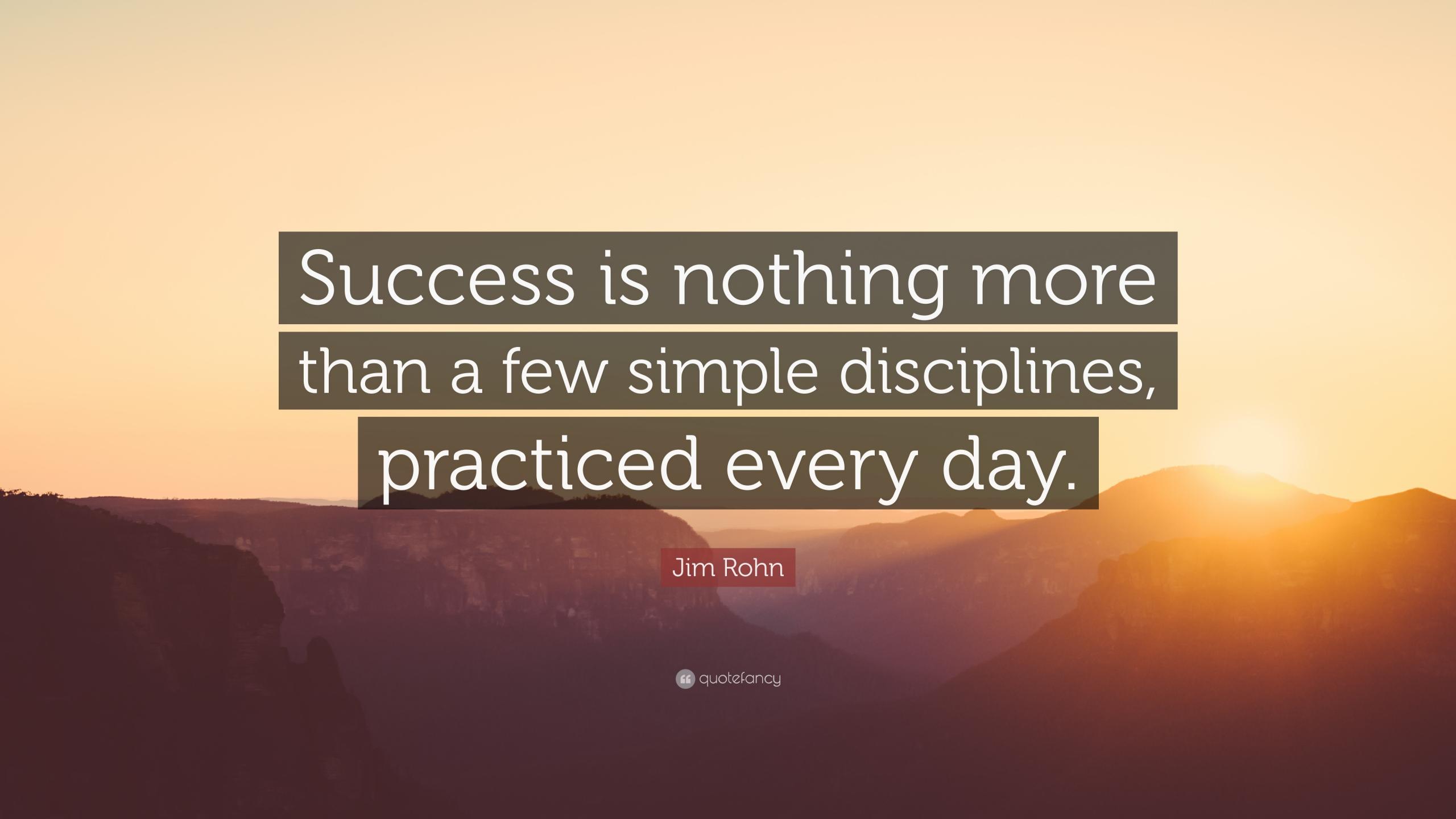


5

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.

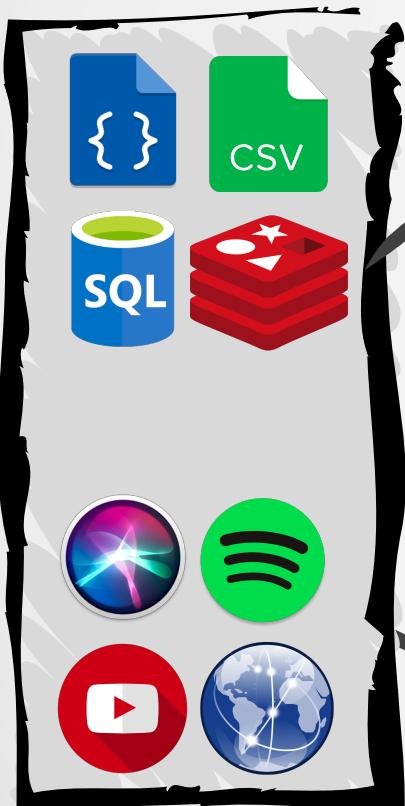
Jim Rohn

Lambda Architecture – Cloud Agnostic & Simplified



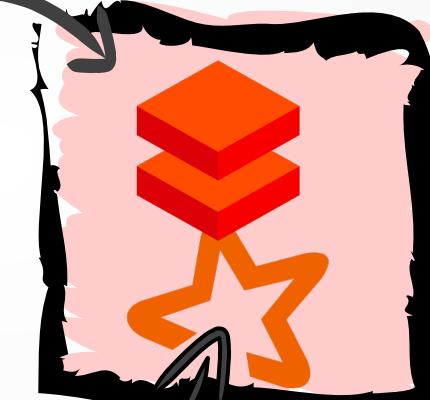
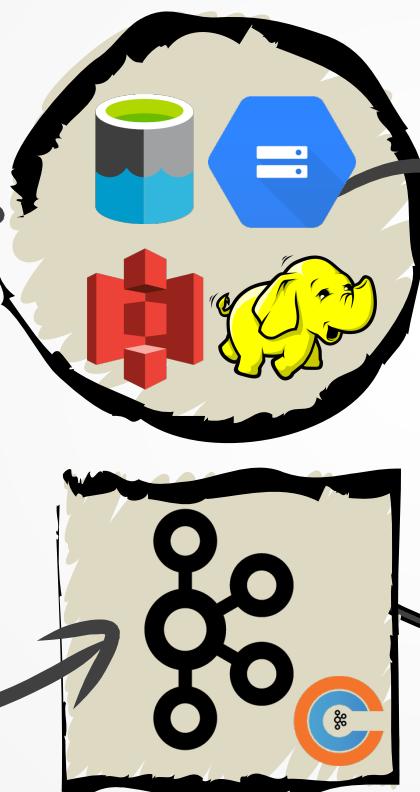
Data Source

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



Speed-Layer

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

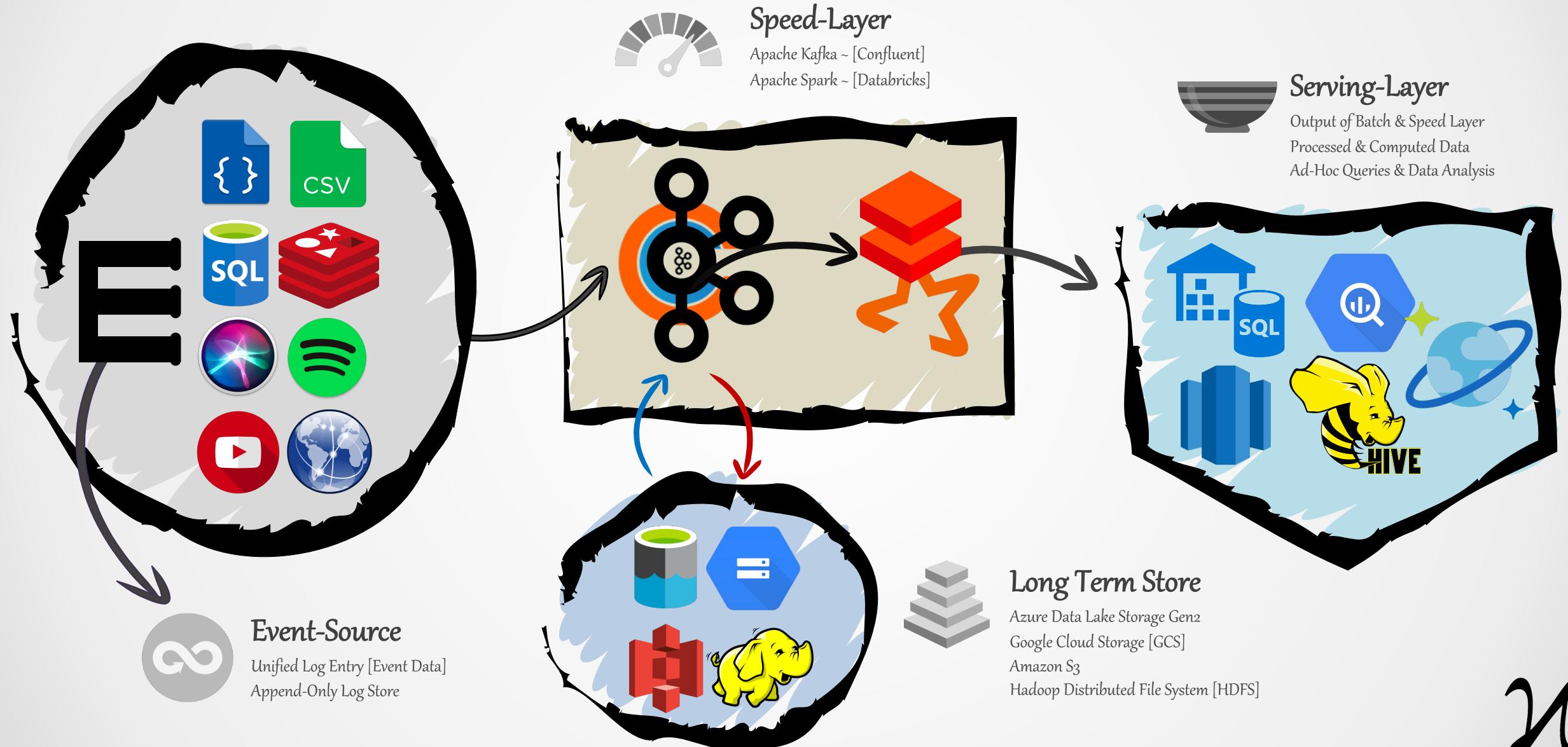


Serving-Layer

Azure SQL Data Warehouse
Amazon Redshift
Google BigQuery
Apache Hive
CosmosDB



Kappa Architecture – Cloud Agnostic & Simplified

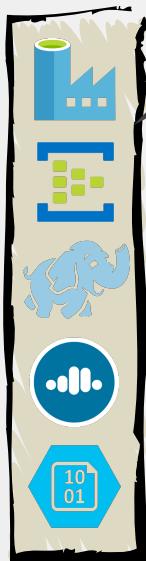


Data Pipeline on Microsoft Azure



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



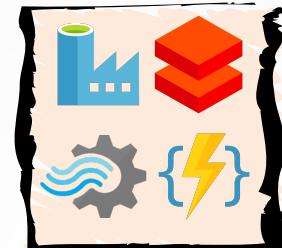
Data Exploration

Azure Data Explorer



Data Processing

Azure Data Factory ~ [Data Flow]
Azure Databricks
Azure Stream Analytics
Azure Functions



Data Serving

HDInsight ~ [Apache Hive]
HDInsight ~ [Interactive Query]
Azure CosmosDB
Azure Synapse Analytics
Azure Analysis Services



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

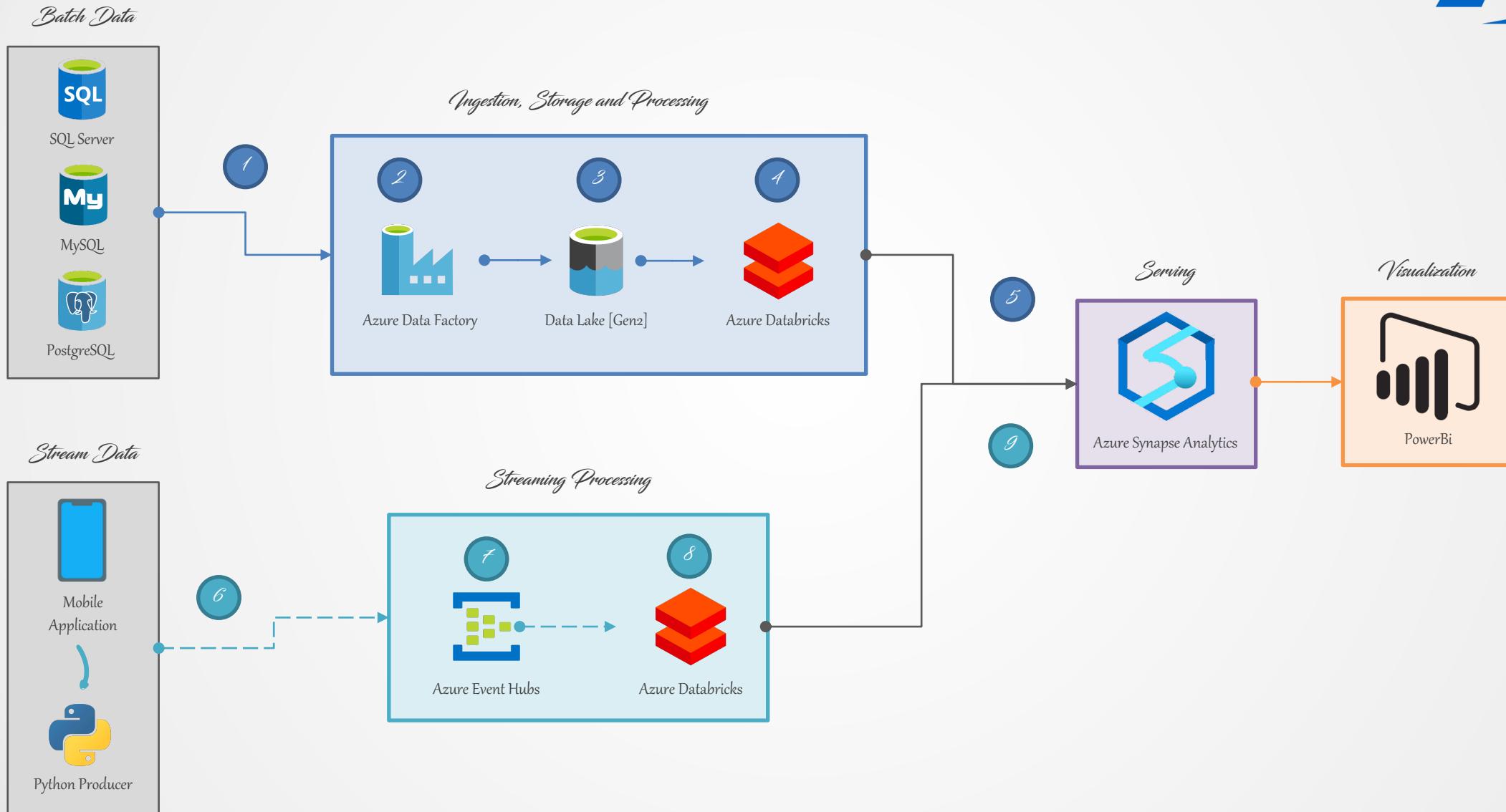


Monitoring

Azure Monitor



Data Pipeline – Lambda Architecture on Azure



Data Pipeline on Google Cloud Platform



Data Ingestion

Google Cloud Pub/Sub
Confluent Cloud
Google Cloud Storage [GCS]
Google Cloud Data Fusion



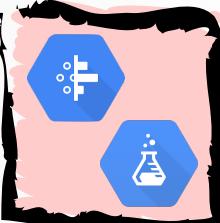
Data Storage

Google Cloud Storage [GCS]



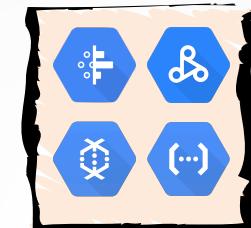
Data Exploration

Google Cloud DataPrep
Google Cloud DataLab



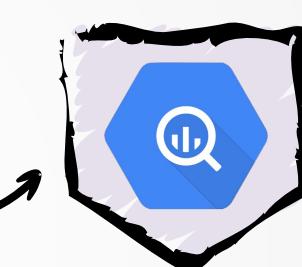
Data Processing

Google Cloud DataPrep
Google Cloud DataProc
Google Cloud DataFlow
Google Cloud Functions



Data Serving

Google BigQuery



Data Viz

Data Studio
PowerBi
Tableau
Qlik
Metabase



Shared Resources

Shared Among Pipeline



Data Discovery

Google Cloud Data Catalog



RDBMS

Google Cloud SQL
Google Cloud Spanner



NoSQL

Google Cloud BigTable
Google Cloud Firestore
Google Cloud MemoryStore



Data Orchestration

Google Cloud Composer



Monitoring

Google Cloud Stackdriver



Data Pipeline on Amazon AWS

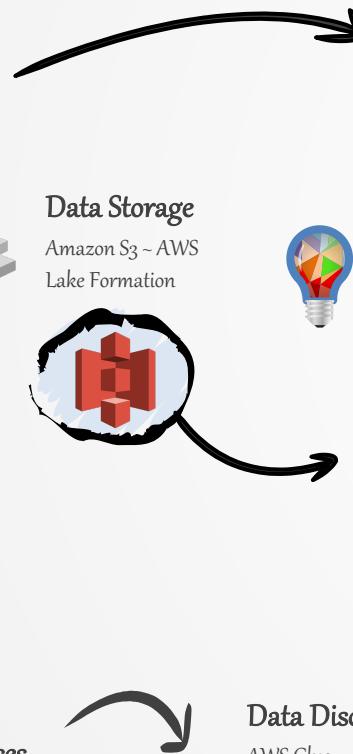


Data Ingestion

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

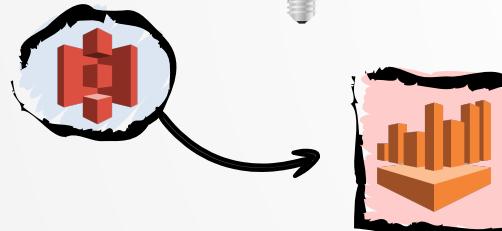


Shared Resources
Shared Among Pipeline



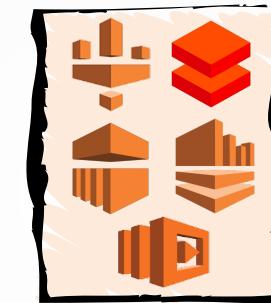
Data Storage
Amazon S3 ~ AWS Lake Formation

Data Exploration
Amazon Athena



Data Processing

AWS Glue
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



Data Discovery
AWS Glue



RDBMS
Amazon RDS



NoSQL

Amazon DynamoDB
Amazon Neptune
Amazon ElastiCache



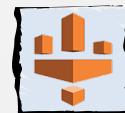
Search

Amazon CloudSearch



Data Orchestration

AWS Glue



Monitoring

Amazon CloudWatch



Use-Case – The Layered Big Data Architecture



Data Source

- Application
- CSV-Parquet

Ingestion Layer

- Apache Kafka
- Azure Blob Storage

Processing Layer

- Azure Databricks

Storage Layer

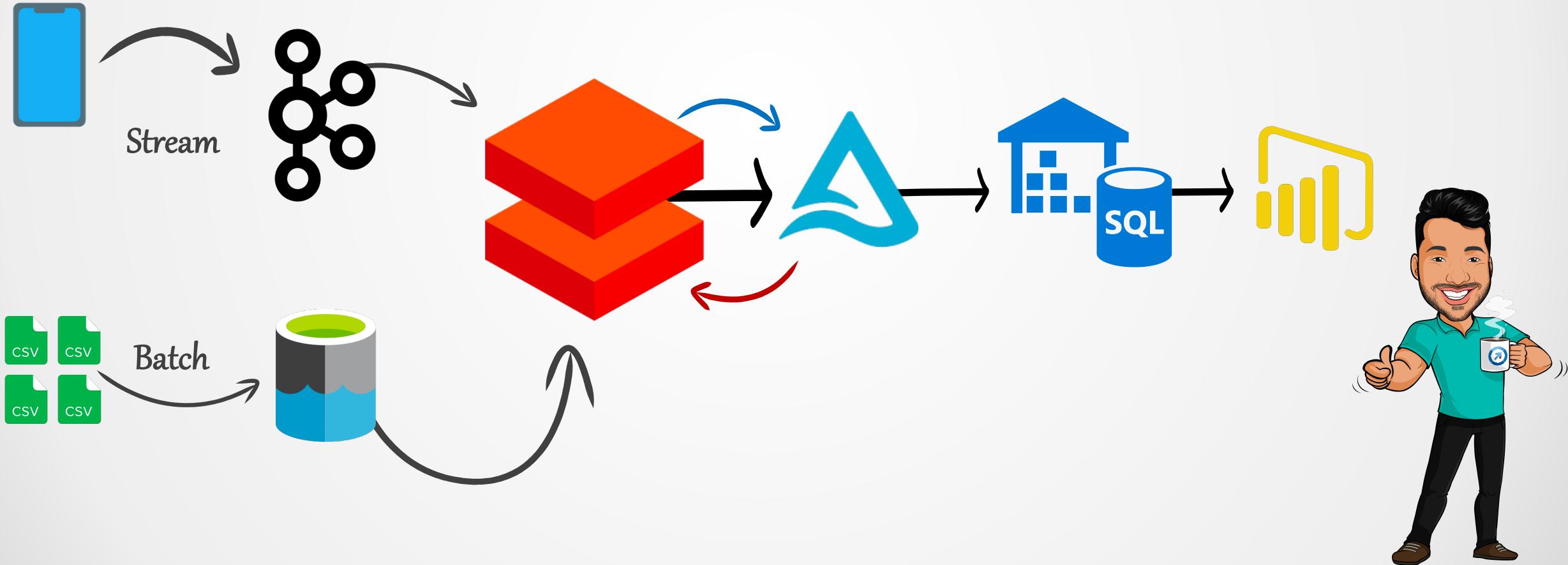
- Delta Lake

Serving Layer

- Azure Synapse Analytics

Presentation Layer

- PowerBi



The Data Engineering Pipeline Project

Layered Big Data Architecture

- Data Sources
- Ingestion
- Processing
- Storage
- Serving
- Presentation



Analyzing Data Sources & Periodicity



Apache Kafka & Azure Blob Storage



Data Processing & Storage Techniques



Serving & Visualizing Data



A wide-angle photograph of a mountain range during sunset or sunrise. The sky is a gradient of orange, yellow, and blue. In the foreground, there's a dark, snow-dusted forest. Behind it, a range of mountains is partially covered in snow, with a prominent peak in the center. The base of the mountains is obscured by a layer of low-hanging clouds or fog. A small, snow-covered town or valley is visible at the bottom of the frame.

Action is the foundational
key to all success.

Pablo Picasso

The Data Architecture Landscape



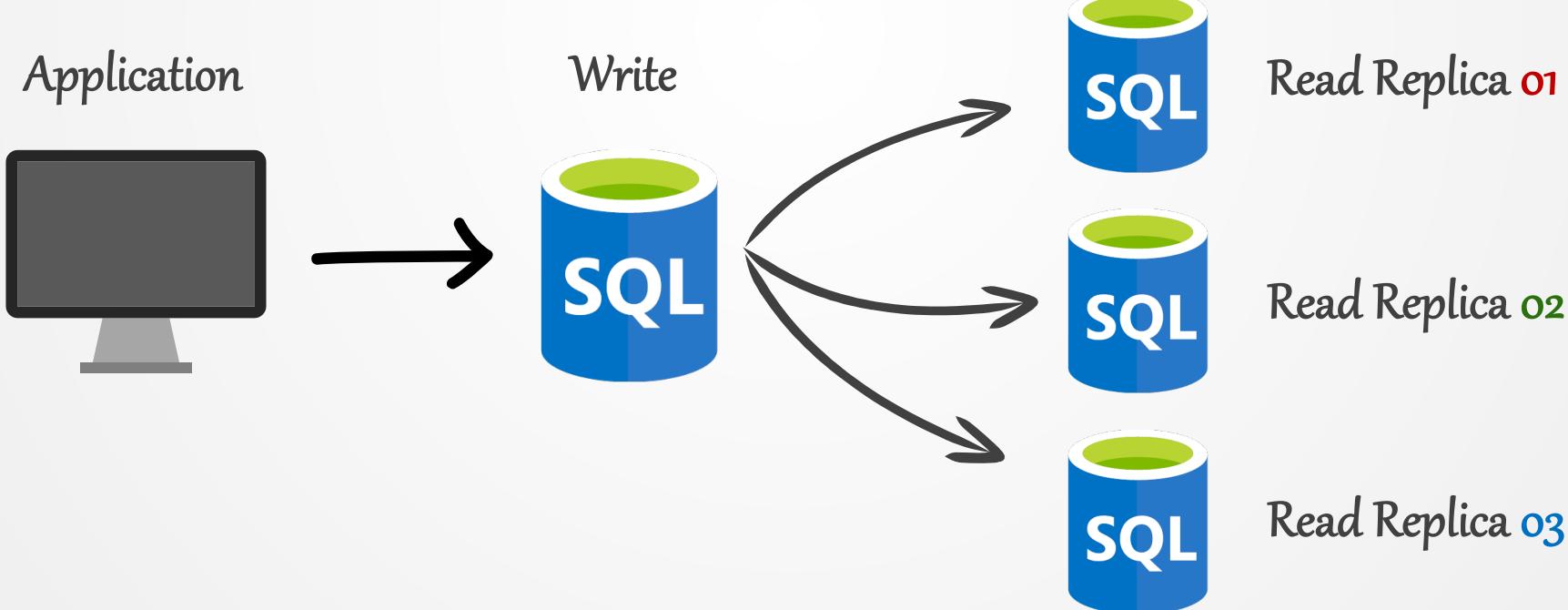
Transactional [OLTP] - 1970

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



Known 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



Known 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



Known Issues

Expensive
Data Silos
Complex

OLTP System



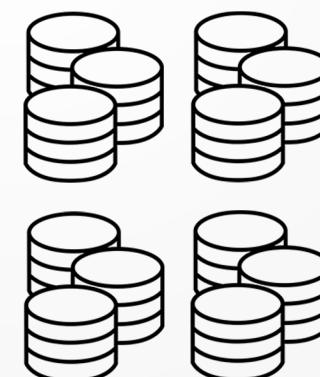
ETL

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



Known Issues

Expensive
Data Silos
Complex

OLTP System



ETL

Extract | Transform | Load



Operational Data Store [ODS]

Departmental Business Decision



ETL

Extract | Transform | Load



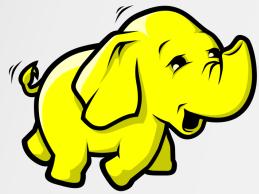
Data Warehouse [Dw]

Ralph Kimball [Bottom-Up]



Volume

The Data Architecture Landscape



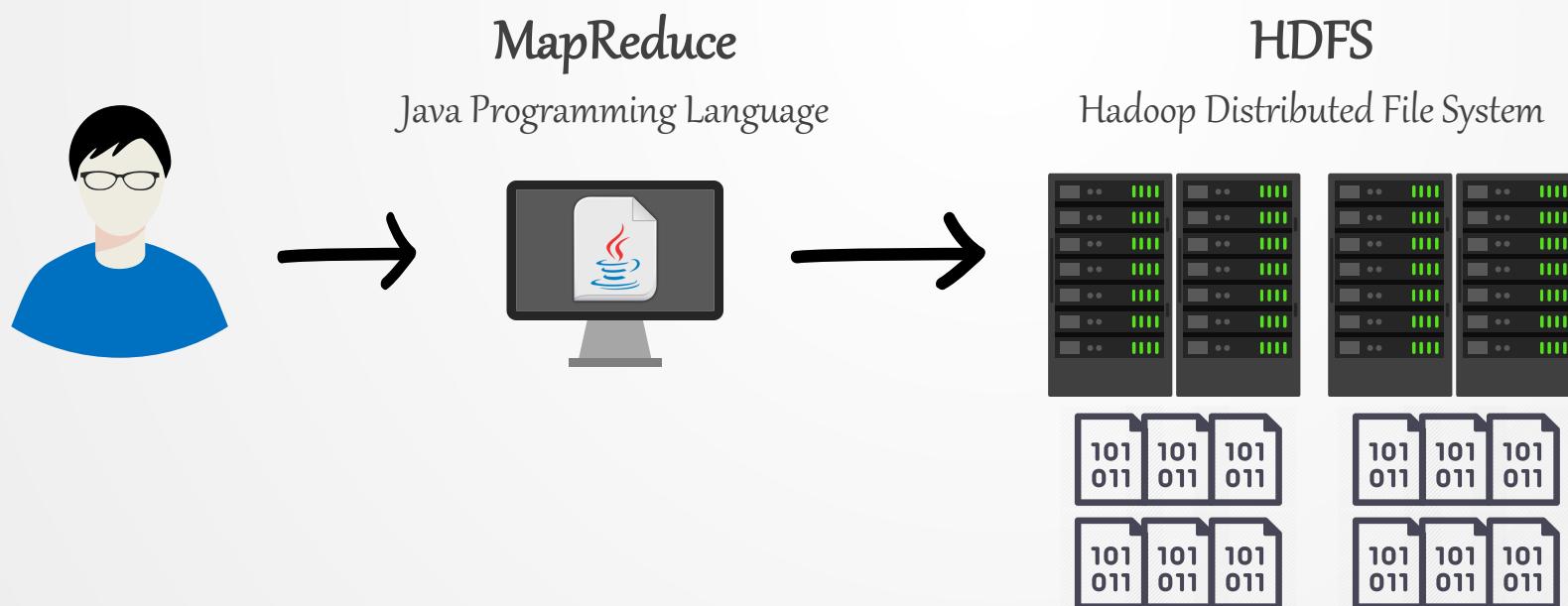
Hadoop – 2006

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



Known Issues

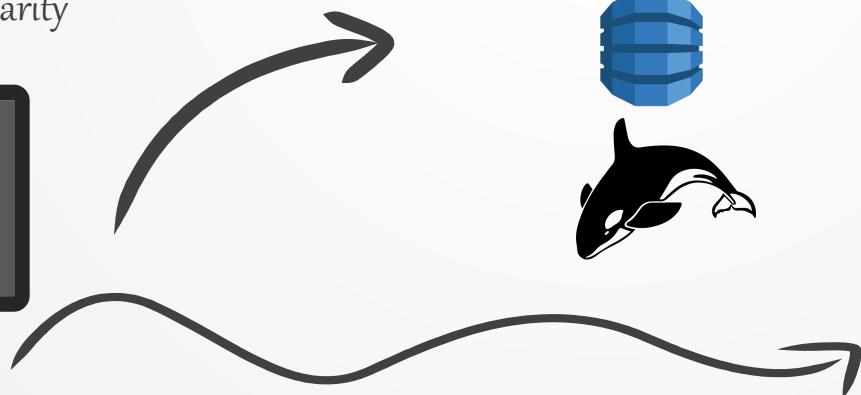
Data Consistency

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



Known Issues

Data Consistency

Lack of Standardization

Scalability [Auto-Capabilities]

Micro-Services Application

Serve Only One Purpose

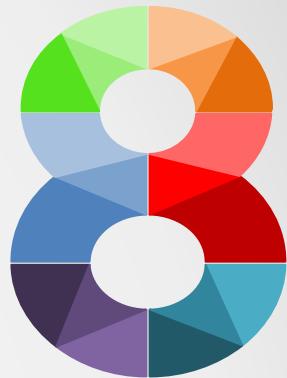


NoSQL

DynamoDB | HBase | MongoDB | Neo4j



The Data Architecture Landscape



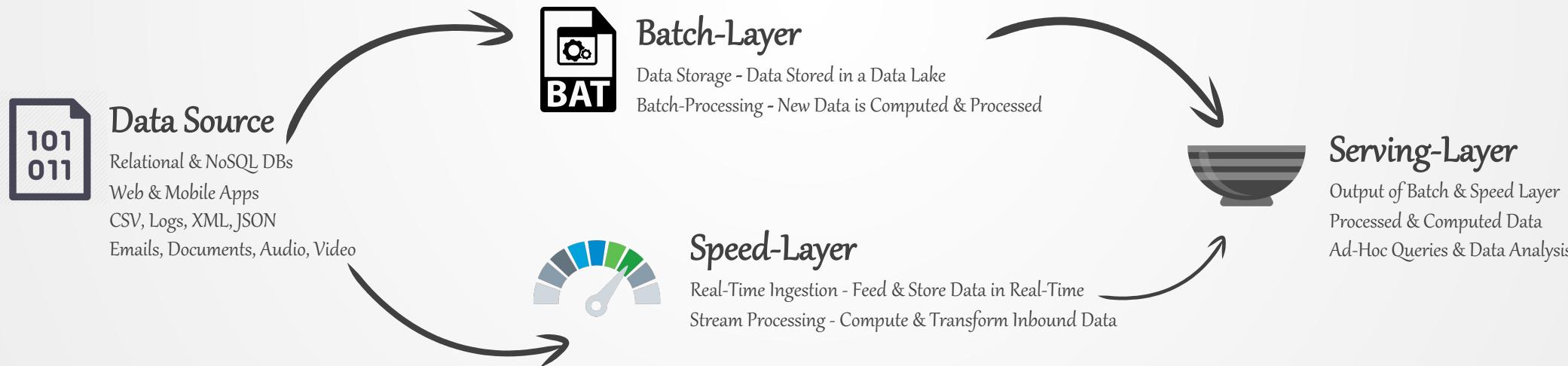
Lambda Architecture – 2013 [1.0]

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



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



κ

Kappa Architecture with *Steroids* – 2019 [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



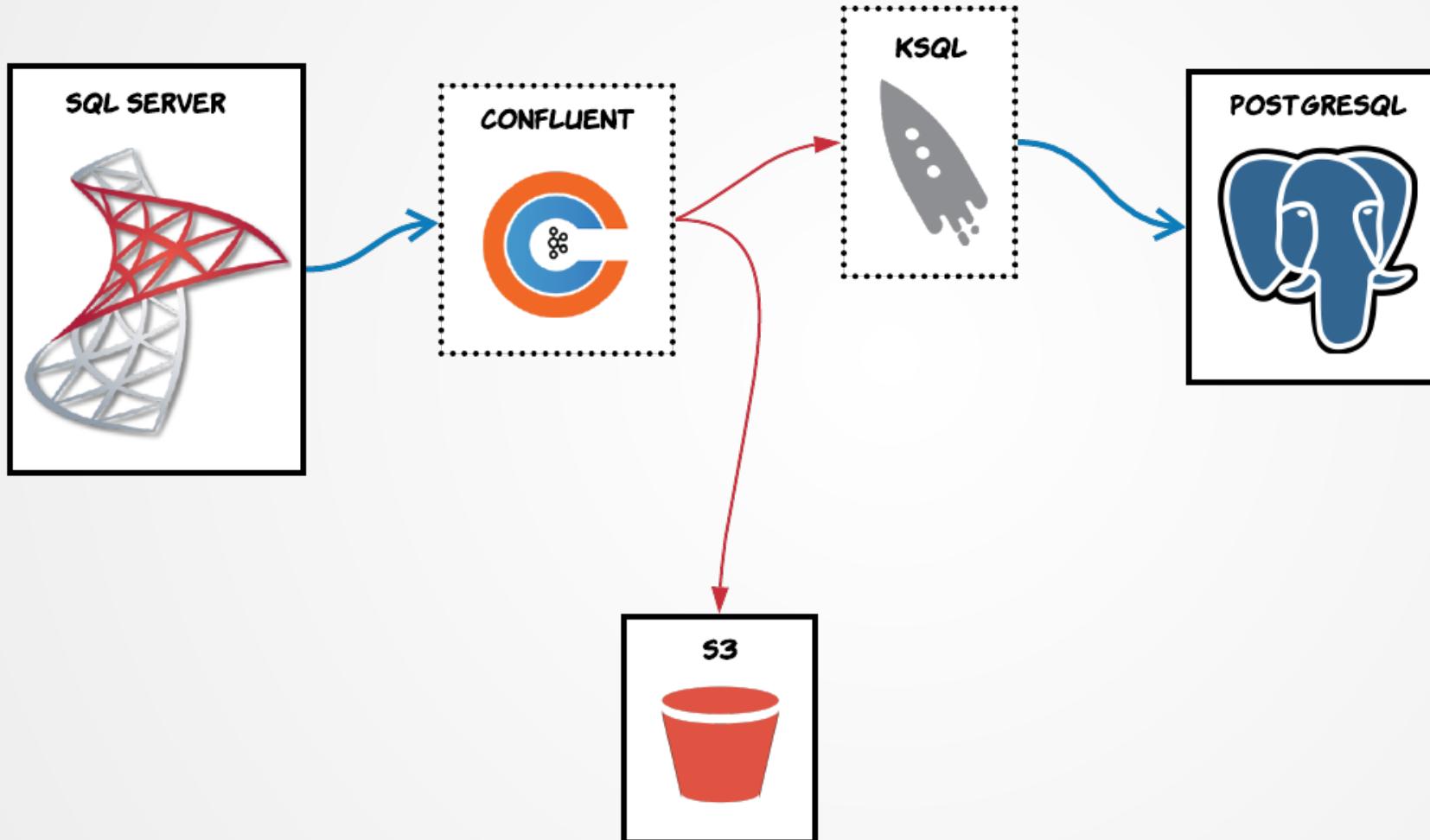
Ingestion Tables

Refined Tables

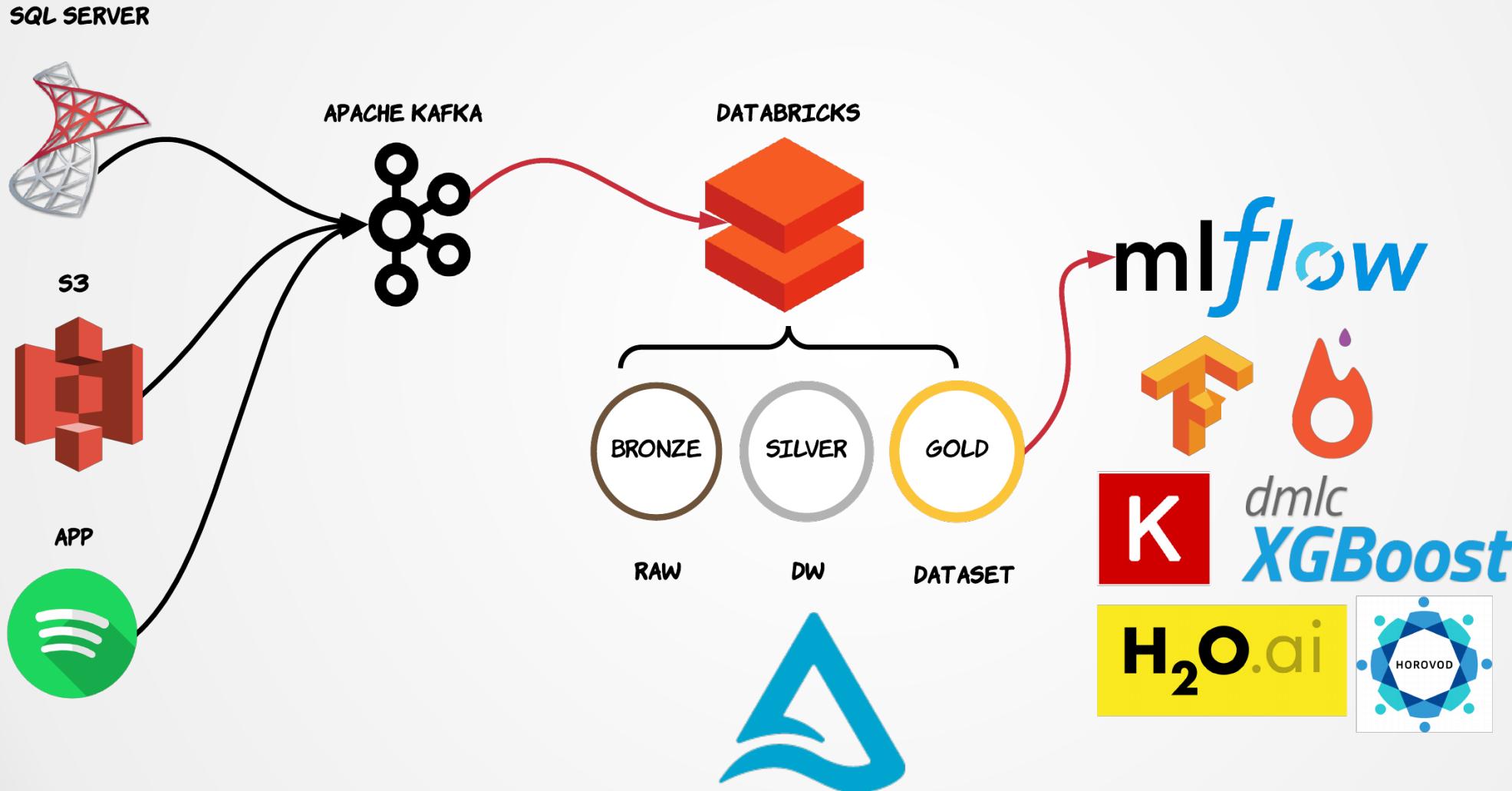
Agg Data Store

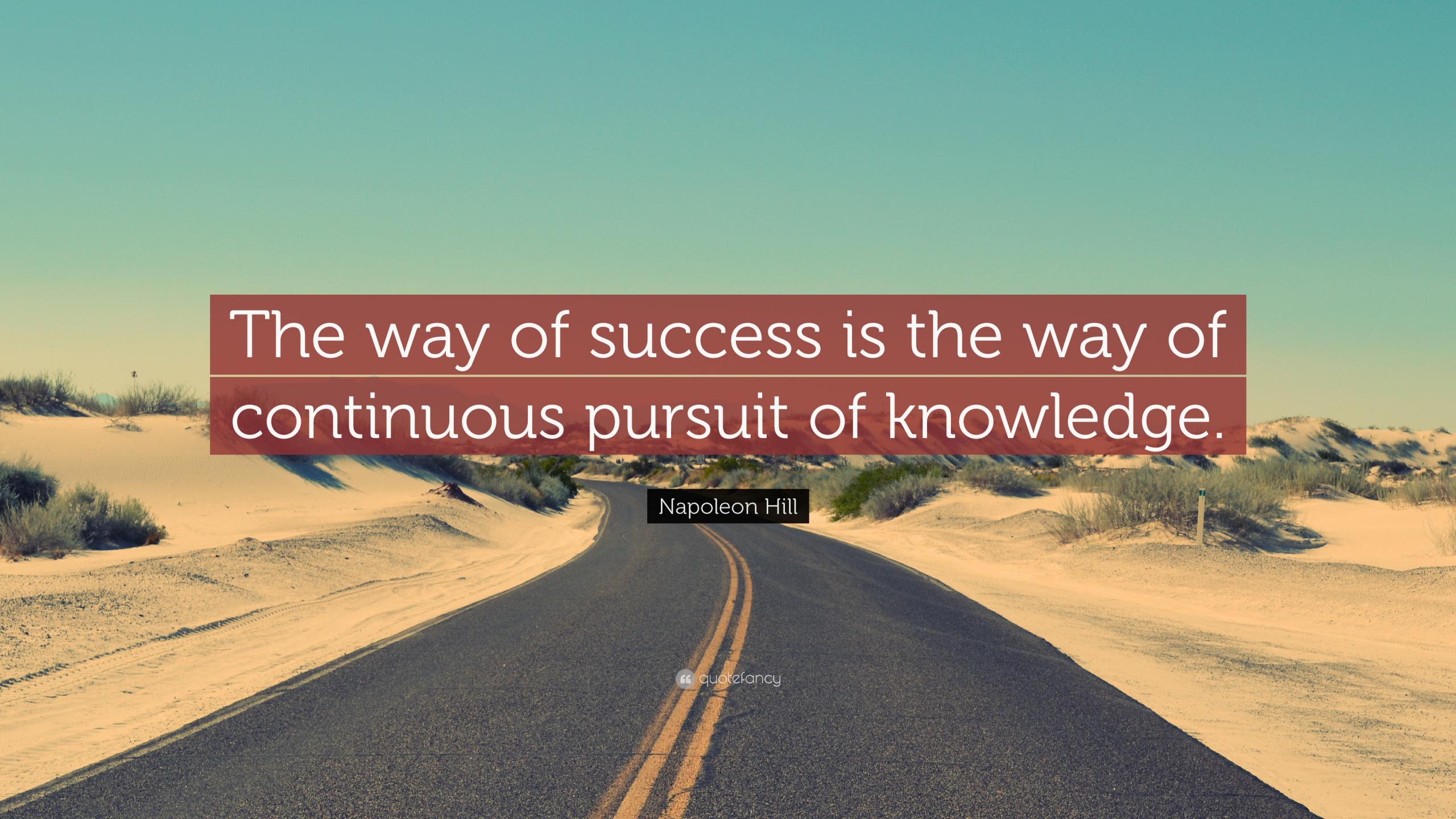


The Data Architecture Landscape – The State of Art



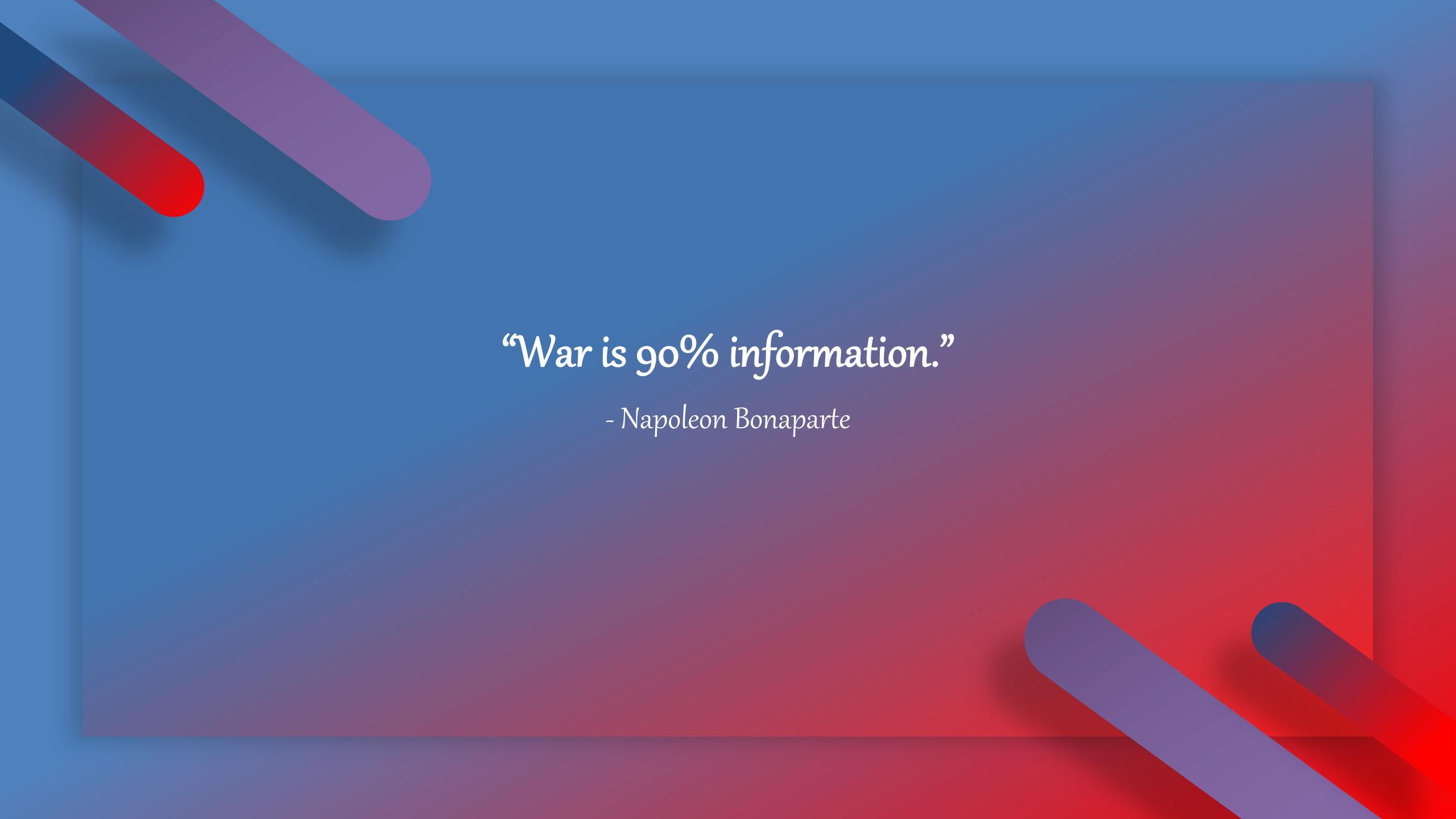
The Data Architecture Landscape – The State of Art



A photograph of a paved road with yellow double lines, curving through a desert environment. Sand dunes are on both sides of the road, and some low-lying desert vegetation is visible. The sky is a clear, pale blue.

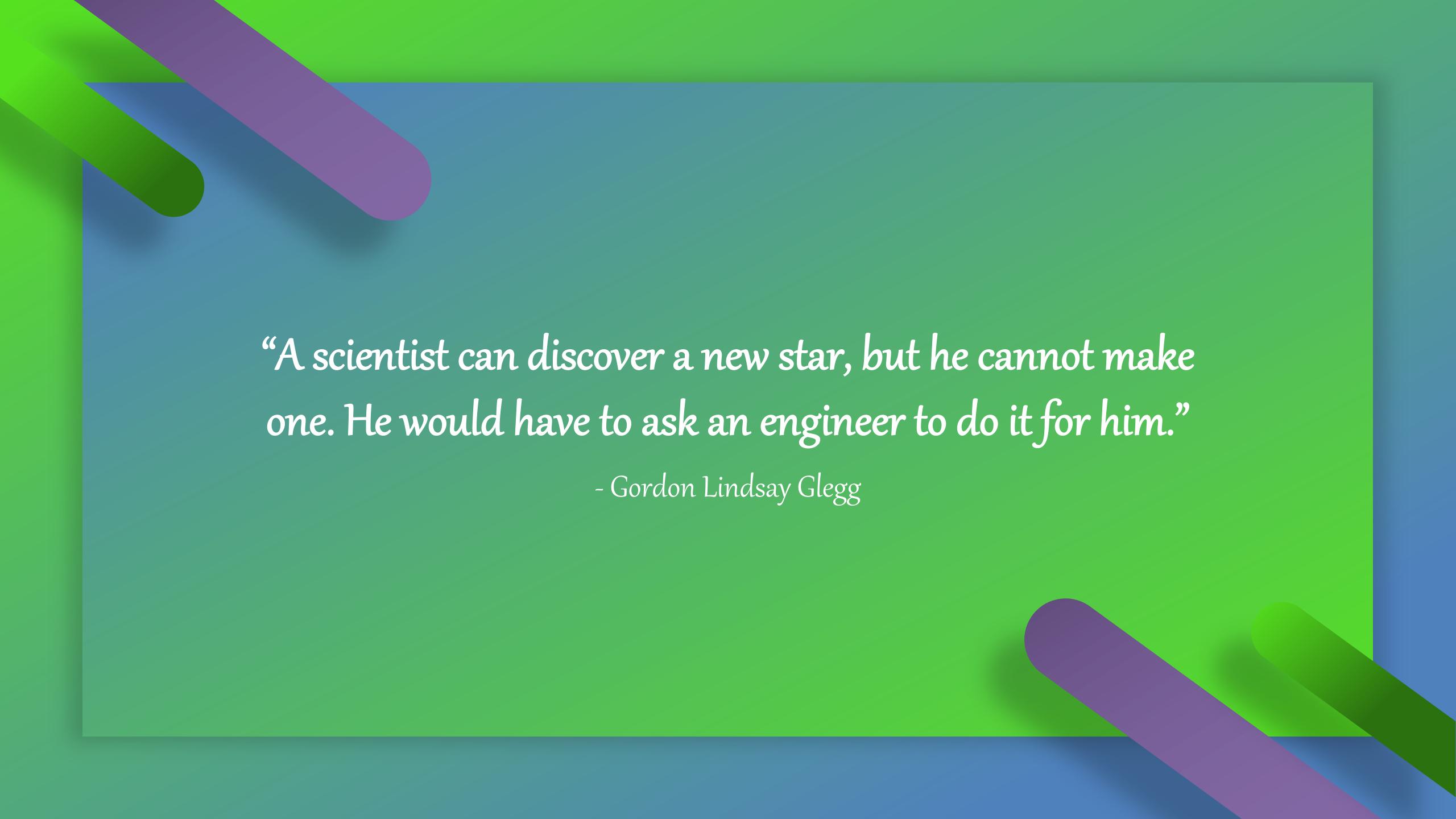
The way of success is the way of
continuous pursuit of knowledge.

Napoleon Hill



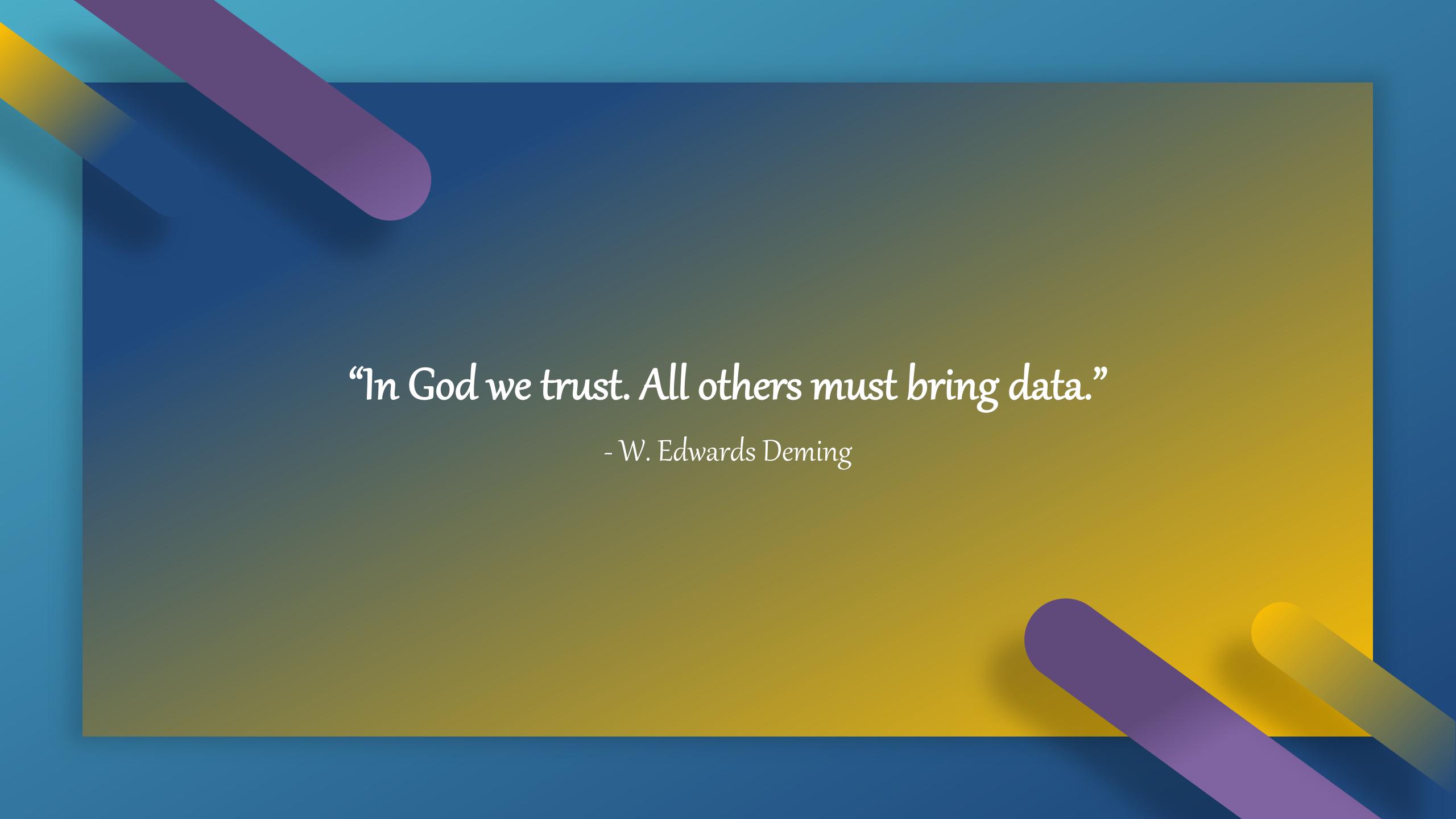
“War is 90% information.”

- 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



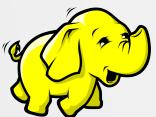
DBMS & NoSQL

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



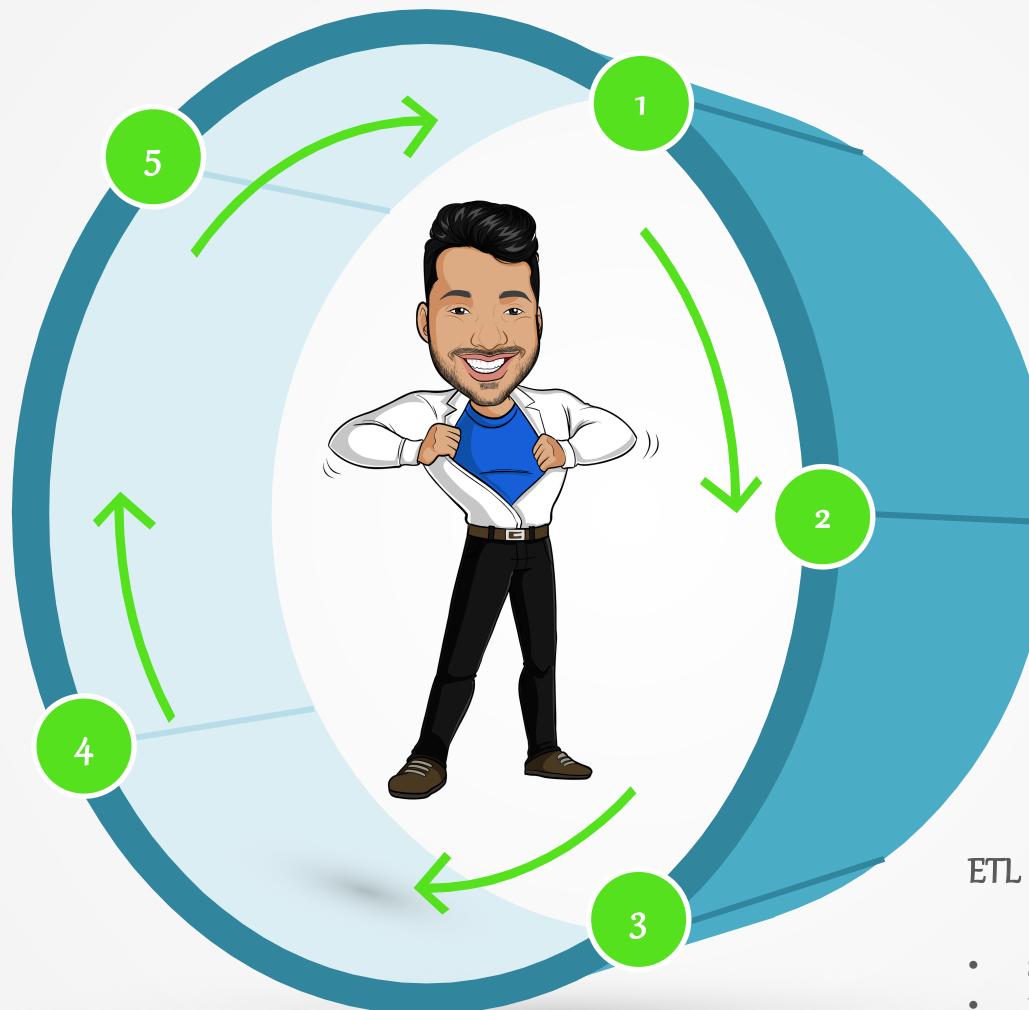
ETL & DW

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



Distributed Systems & Big Data Frameworks

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



Data Pipelines & Cloud Computing

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

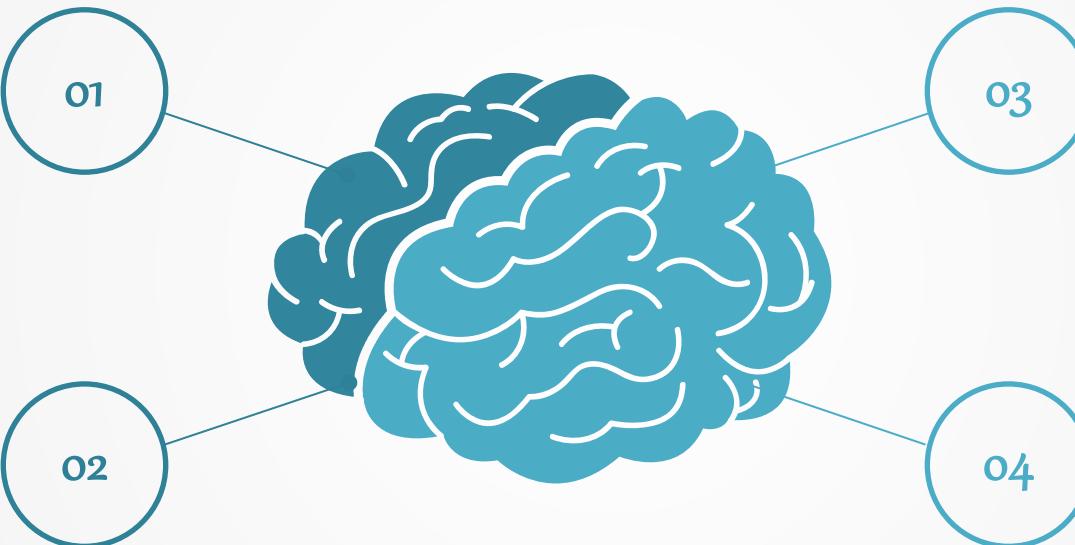
Data Engineer Business Skills

Data Engineer Career - Part 2



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.



Effective Collaboration

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

Intellectual Curiosity

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

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.



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.



CRT020: Databricks Certified Associate Developer for Apache Spark 2.4

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

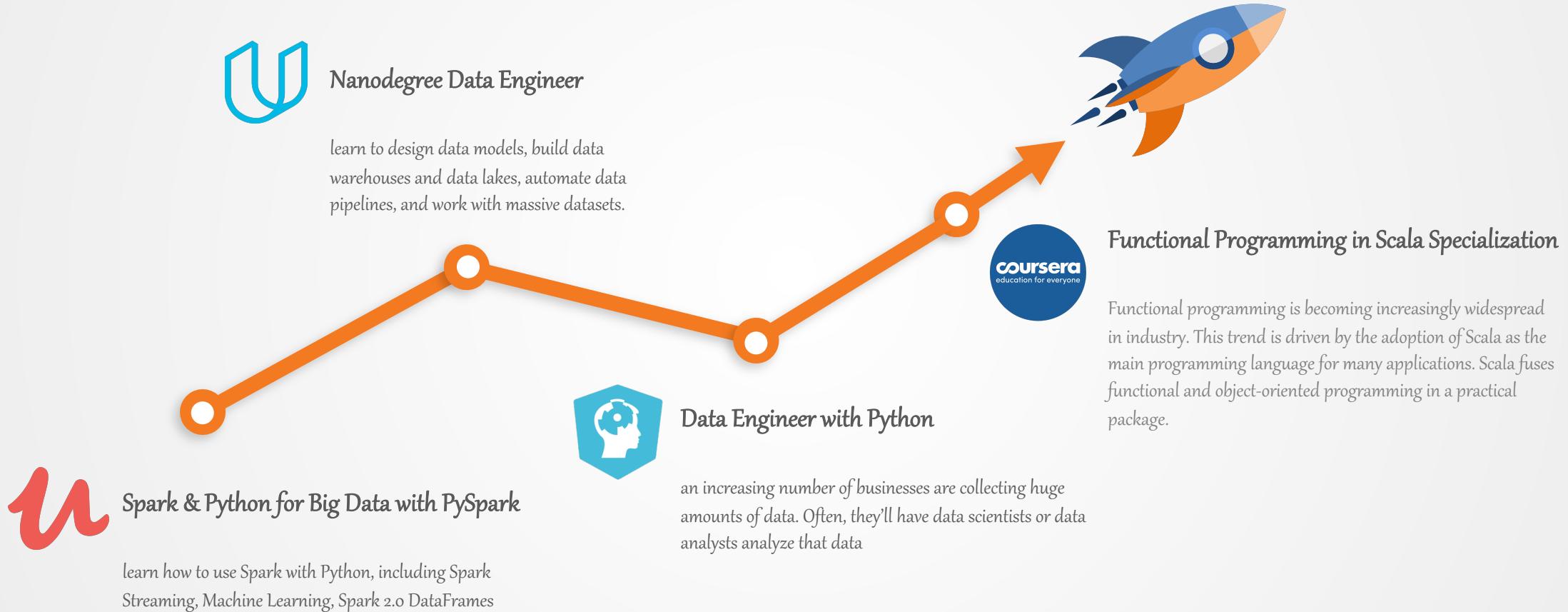


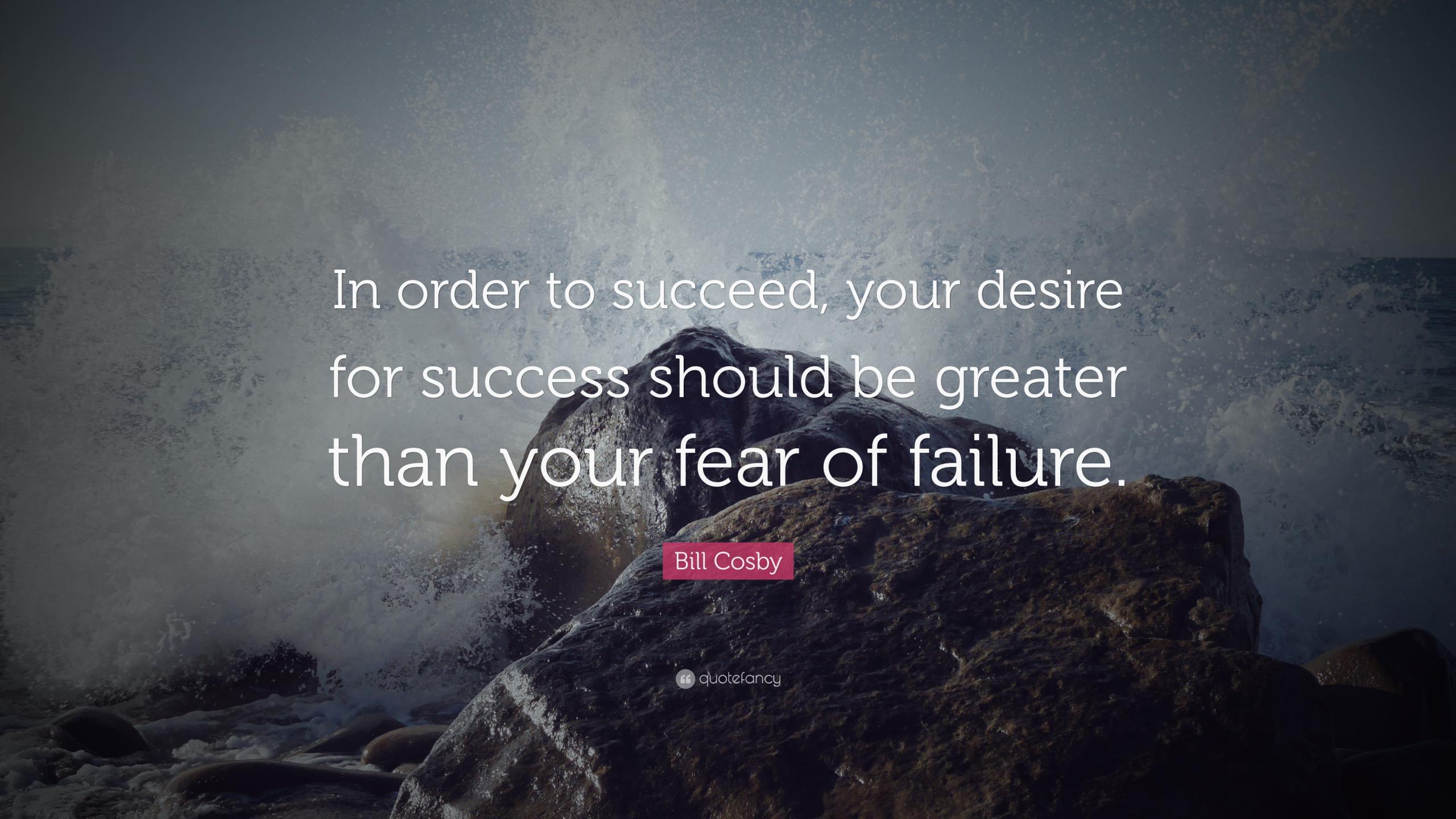
Confluent Certified Developer for Apache Kafka (CCDAK)

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

Data Engineer Study

Data Engineer Career - Part 4



A dark, moody landscape featuring a large, rugged mountain peak in the foreground. The mountain's surface is textured with dark, weathered rock. In the background, a misty, rocky cliff face is visible, partially obscured by low-hanging clouds or fog. The overall atmosphere is dramatic and inspiring.

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

Bill Cosby

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