ONEWAY

SOLUTION

One Way Solution

Patterns & Common Use-Cases

Data Engineering – [Day 5]

LUAN MORENO

CEO & CDO

Data Engineer & Data Platform MVP

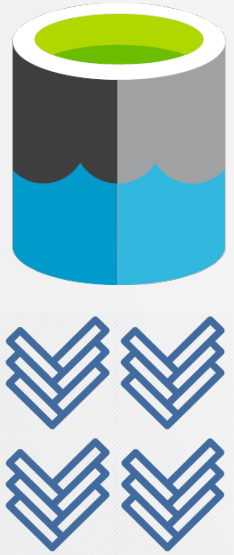
Confluent Certified Developer for Apache Kafka [CCDAK]


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

The Spark Lifecycle

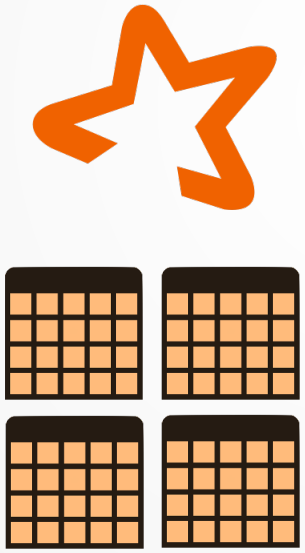
Data Lake

Repository of Raw Data
Without Schema Enforcement



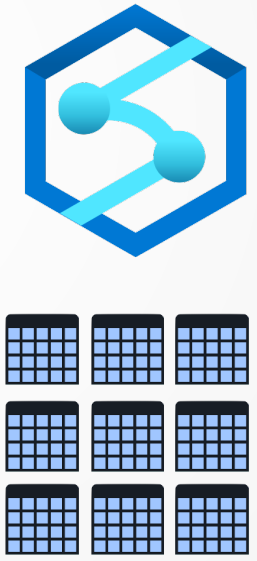
Apache Spark

Distributed Cluster-Computing Framework
Optimized for Memory Computation



Data Warehouse

Analytics Platform for Enterprises
Scalability – Horizontally & Vertically

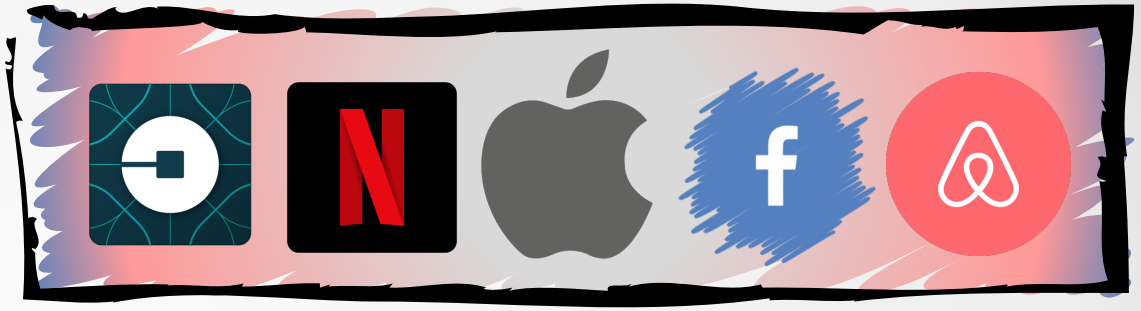


Raw Ingestion

Transformations

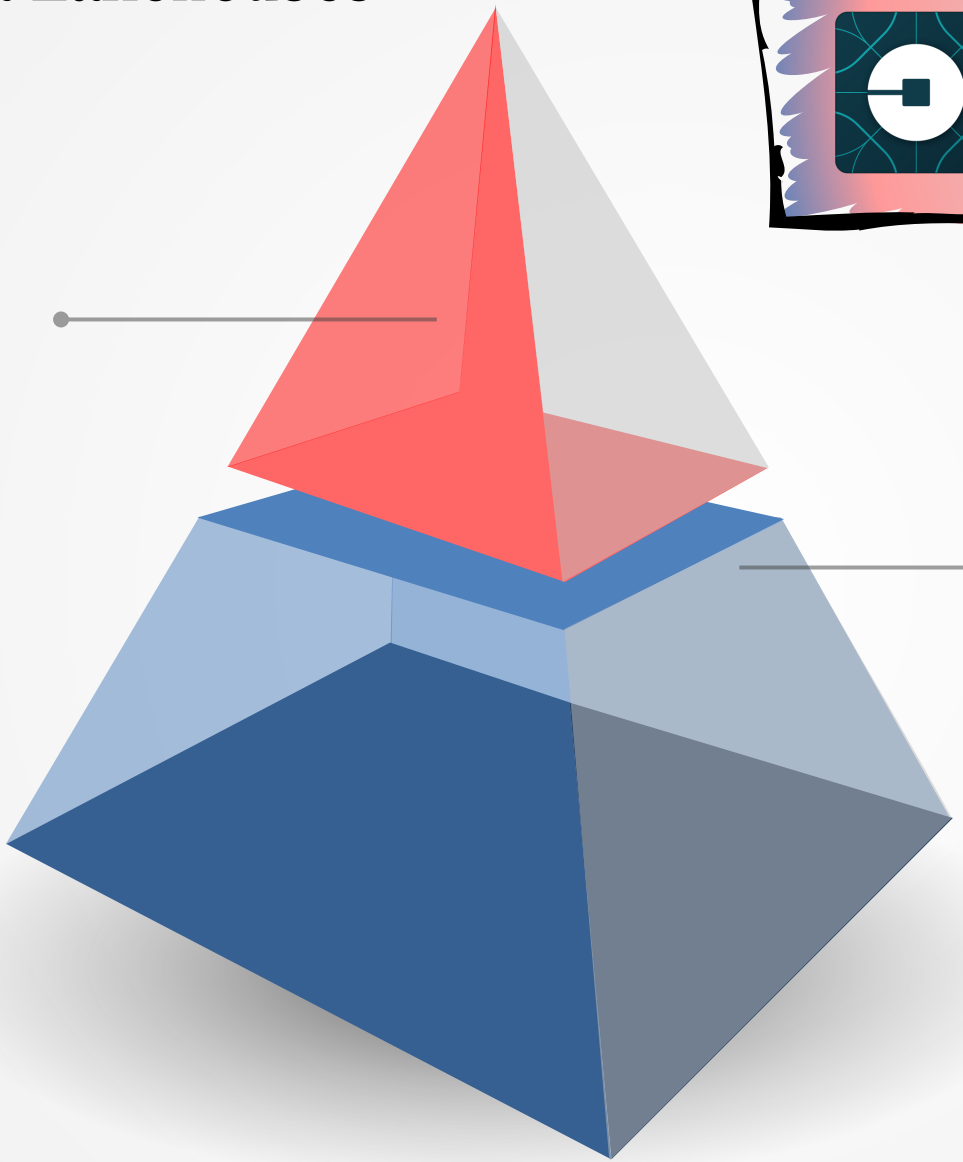
Business-Level

Data Lake vs. Data Lakehouses



Data Lakehouses

Metadata Layers for Data Lakes
New Query Engine Design
High-Performance SQL Execution
Optimized Access of Data



Data Lake

Repository of **Raw** Data
Unsiloed Data
Without Schema Enforcement
Data Swamp & Data Quality Issues

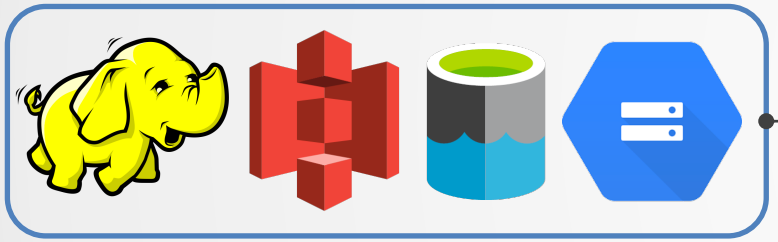
The Delta Architecture

Data Lake

Repository of Raw Data
Without Schema Enforcement

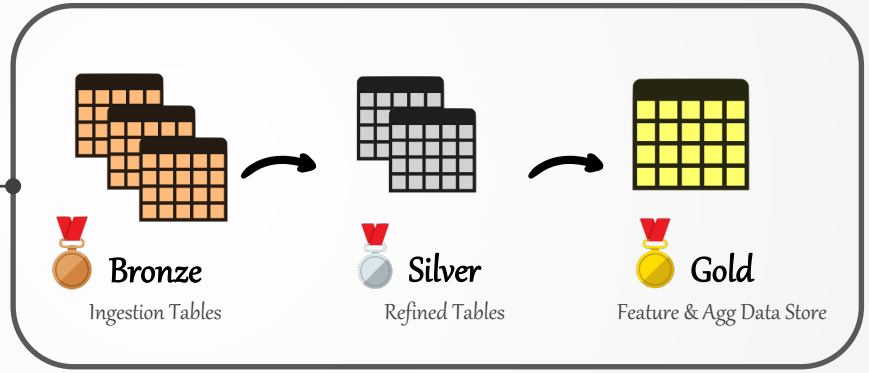
Delta Lake

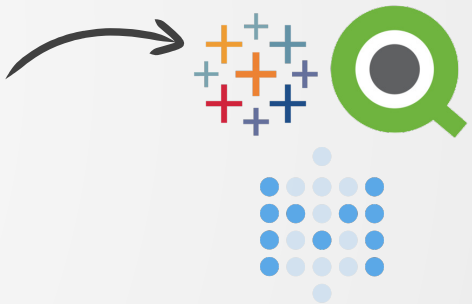
Storage Layer with ACID Transactions
Big Data Workloads

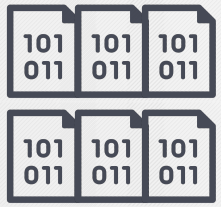


Batch

Stream







Raw Ingestion

Transformations

Business Level


What you focus on grows.
Esther Hicks
quotefancy

Lambda Architecture – Cloud Agnostic & Simplified

Data Source

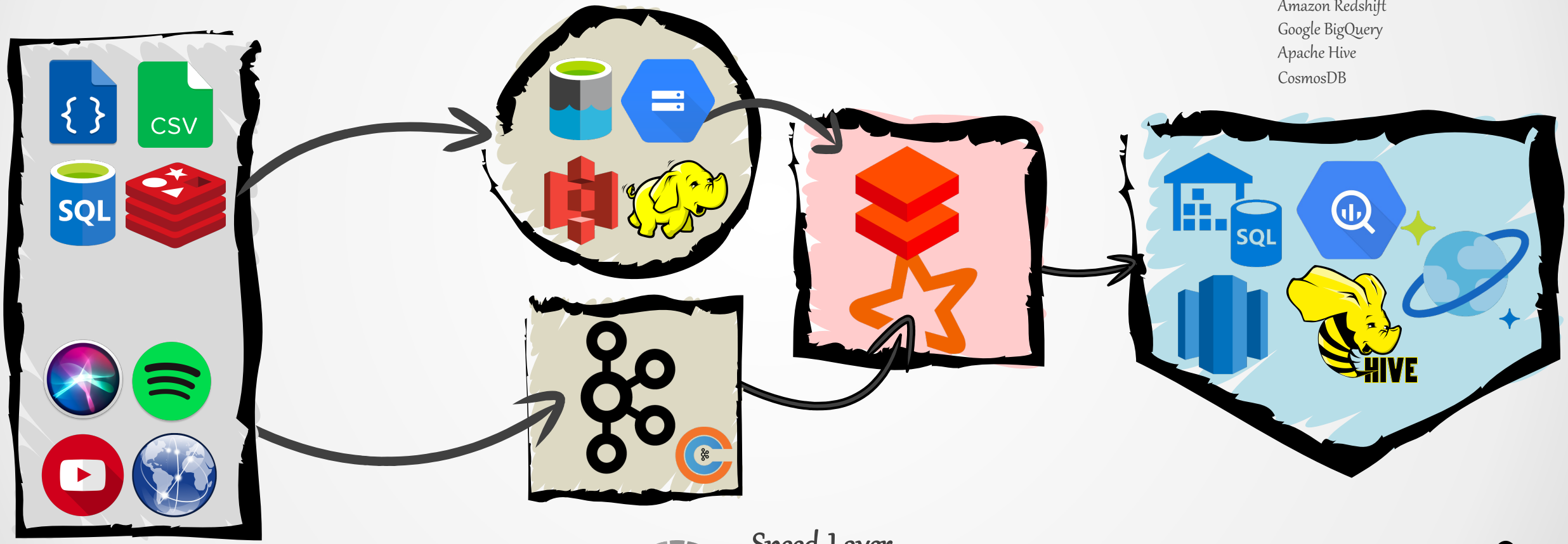
JSON & CSV
SQL Server & Redis
Internet – Siri | Spotify | YouTube

Batch-Layer

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

Serving-Layer

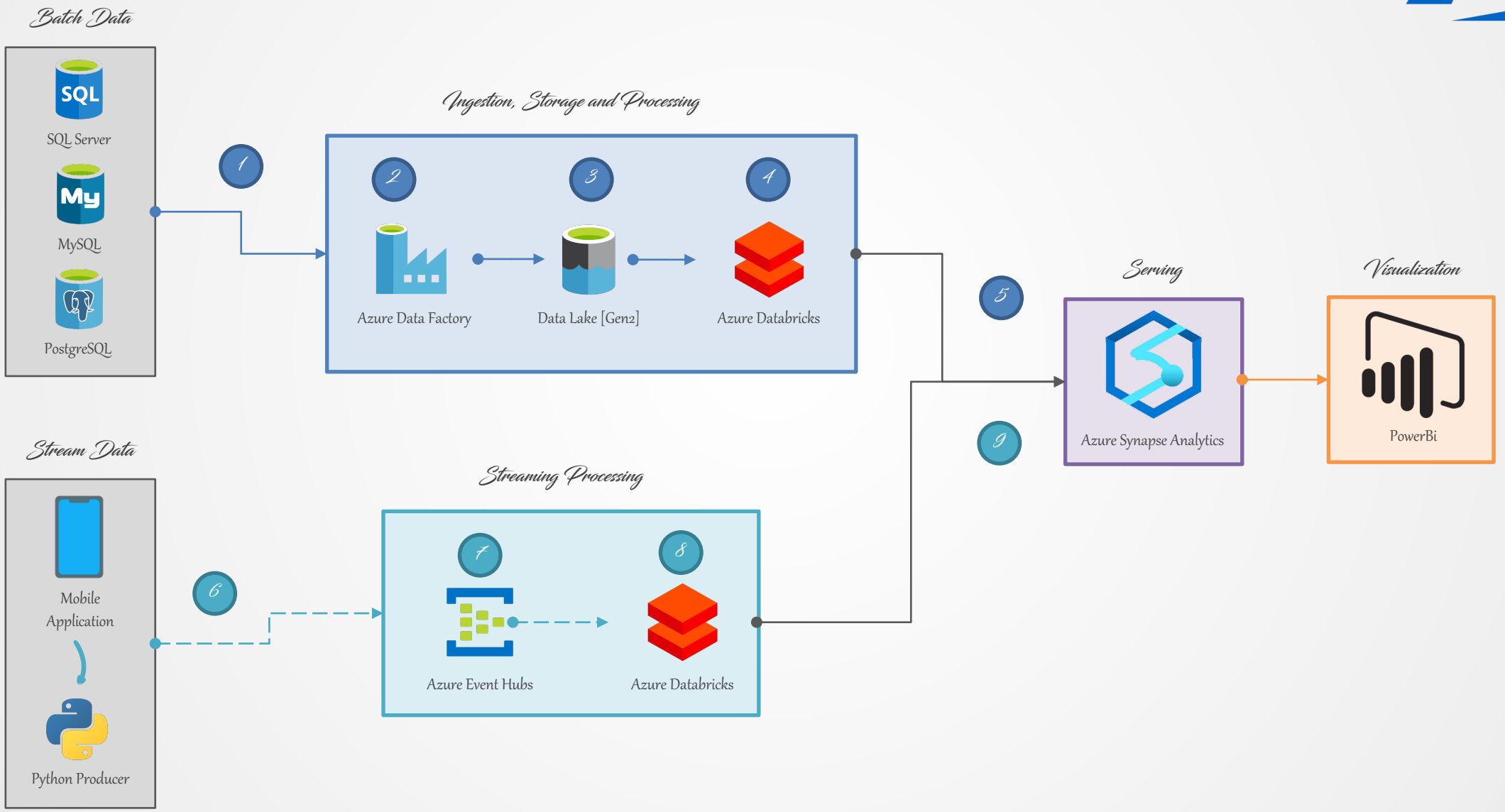
Azure SQL Data Warehouse
Amazon Redshift
Google BigQuery
Apache Hive
CosmosDB



Speed-Layer

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

Lambda Architecture



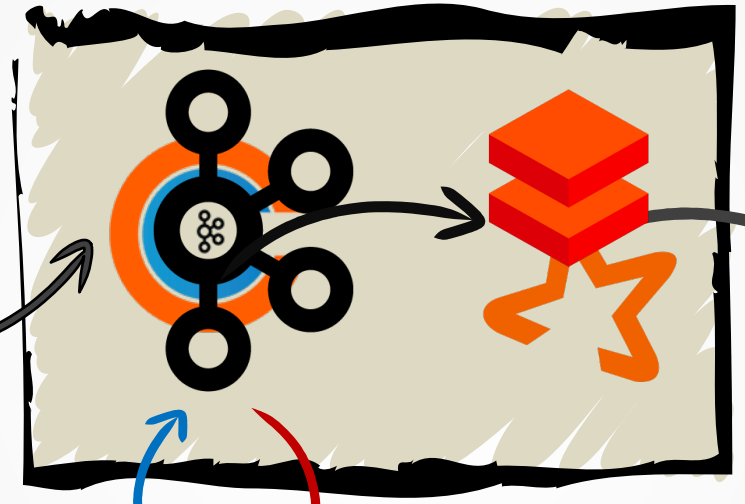
Kappa Architecture – Cloud Agnostic & Simplified

Speed-Layer

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

Serving-Layer

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



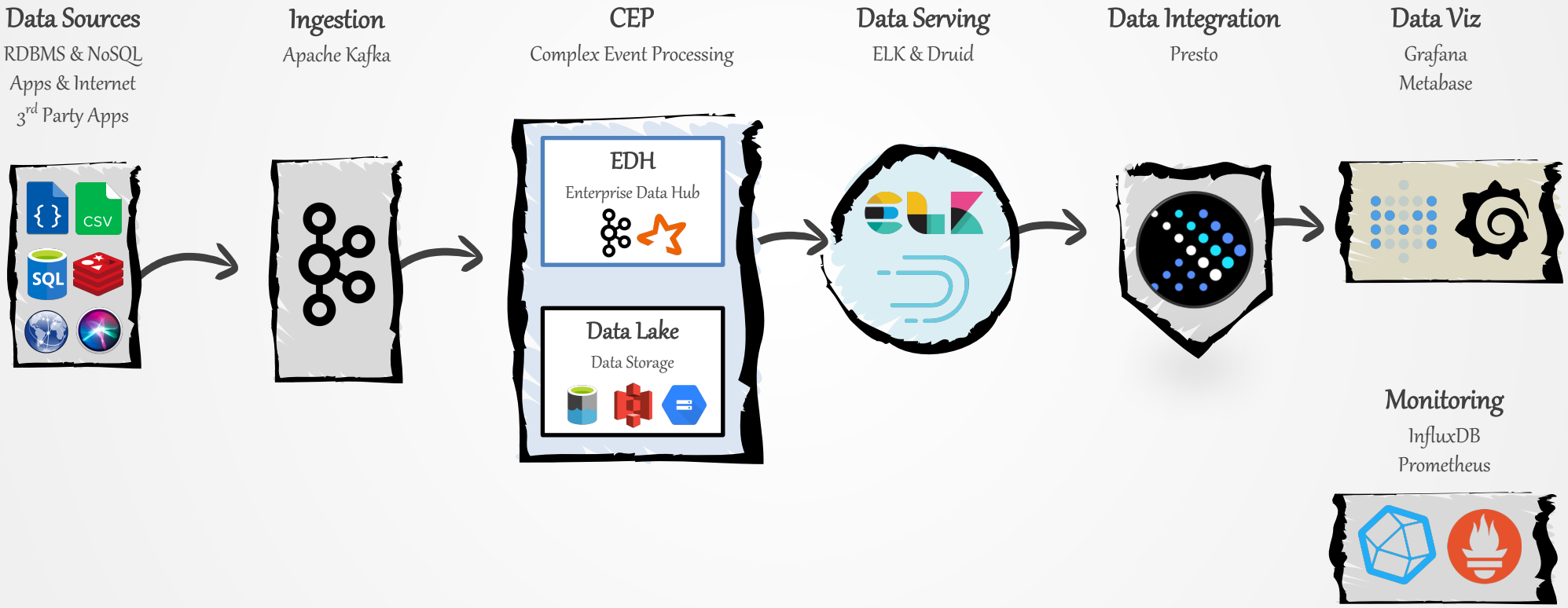
Event-Source

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]

Kappa Architecture

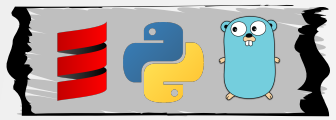


[Orion] - Big Data as a Service

data-driven architecture for microservices & data pipelines. this process allows tighten and seamless integration between applications and data stores for analytics at scale.

Kappa Architecture

Application Ingestion Layer
Kafka Producer API



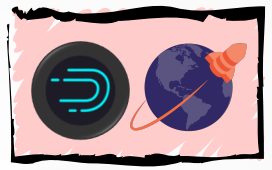
Data Store Ingestion Layer
Kafka Connect Source API

Enterprise Data Hub [EDH]
Apache Kafka [Strimzi]

Deep Storage
MinIO

Processing Layer
KSQLDB, Apache Spark & Faust

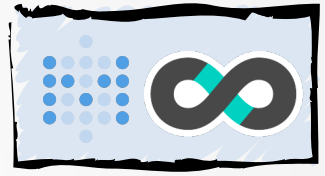
OSS Data Serving
Apache Druid & YugaByteDB



Cloud Data Serving
Redshift, BigQuery, Synapse Analytics

Integration & Exploration Layer
Trino & Zeppelin

Data Visualization
Metabase & Superset



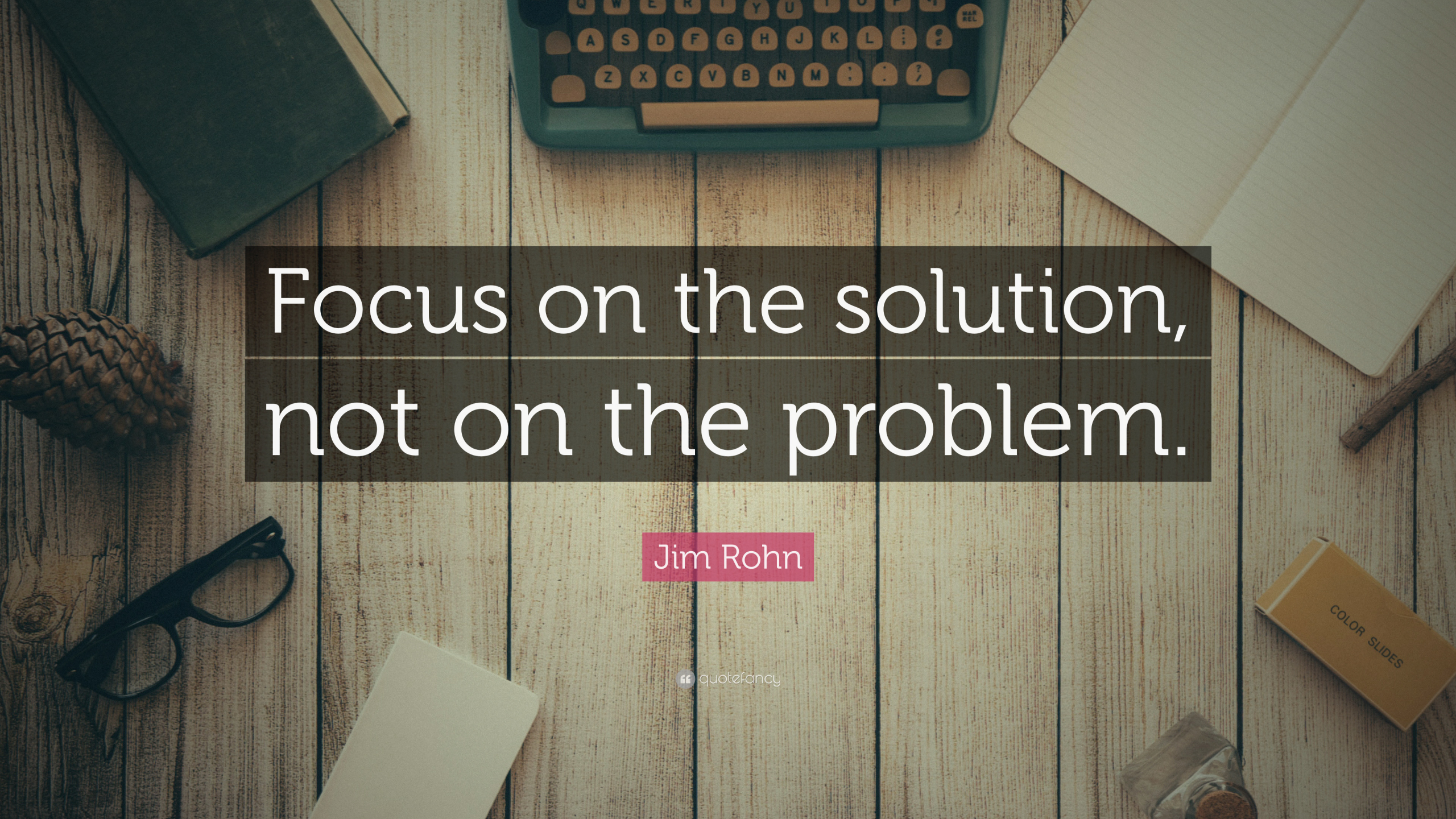
Continuous Delivery
ArgoCD - GitOps

Orchestration Layer
Airflow

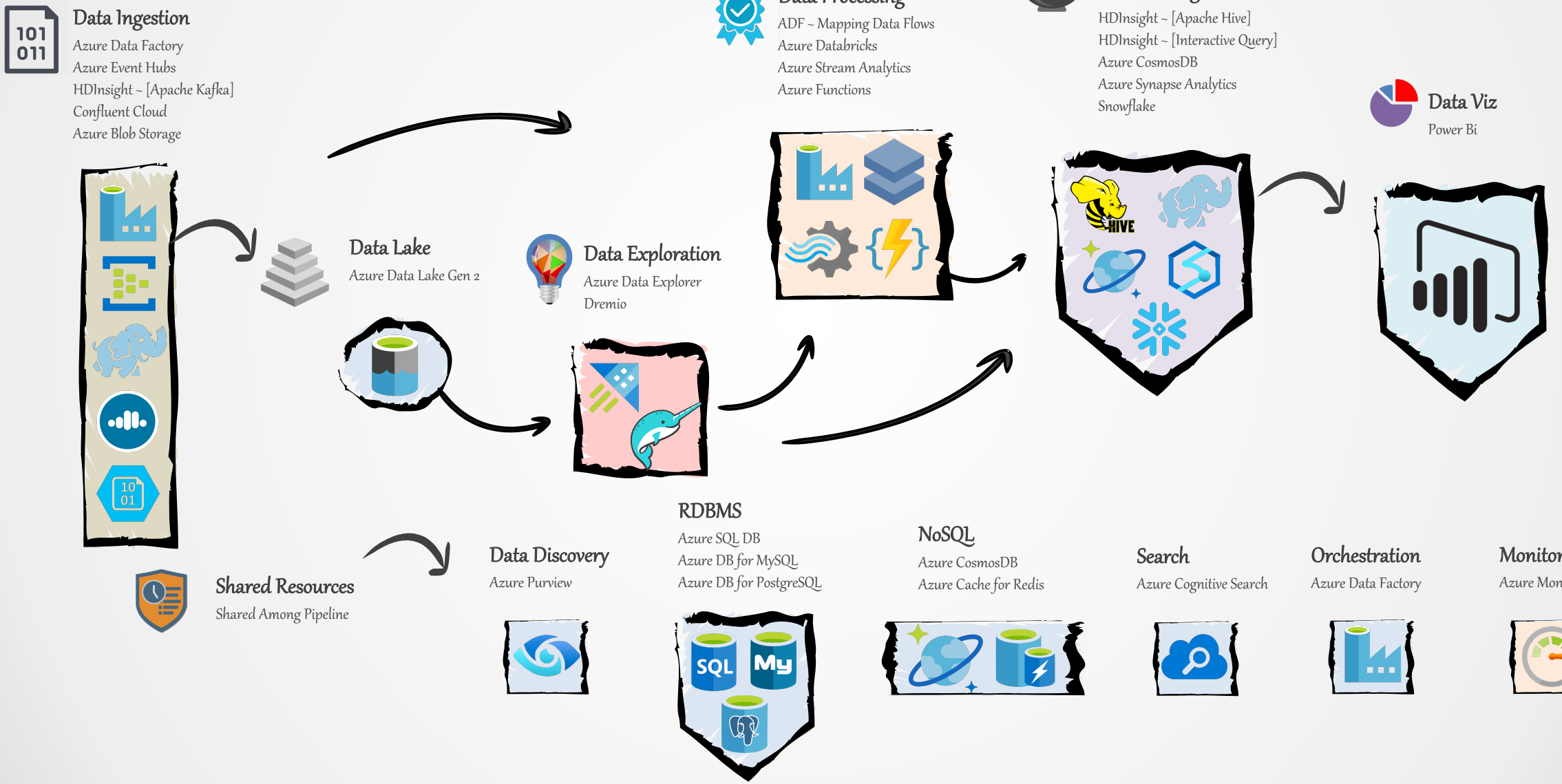
Logging
FileBeat, ElasticSearch & Kibana

Monitoring
Prometheus, Alert Manager & Grafana

Kubernetes
Open-Source Container-Orchestration System


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

Microsoft Azure Big Data Landscape for Data Pipelines



Cost of a Data Pipeline on Microsoft Azure

Data Lake

Raw Repository
Low-Cost Storage

Cost for Azure Blob Storage

- **Region** – EastUS2
- **Tier** – Premium
- **Redundancy** – LRS
- **Capacity** – 1 TB
- **Monthly Cost** – R$ 847

Real-Time Ingestion

Ingesting Events at Scale
Store and Replicate Data Securely

Cost for Azure Event Hubs

- **Region** – EastUS2
- **Tier** – Standard with Capture
- **Ingress** – 1 Million
- **Throughput** – 3 MB
- **Monthly Cost** – R$ 1.567

Data Processing

Data Analytics Platform
Apache Spark Engine

Cost for Synapse Analytics Apache Spark Pools

- **Region** – EastUS2
- **Instance** – Small
- **Spec** – 4 vCPUs & 32 GB of RAM
- **VMs** – 3
- **Monthly Cost** – R$ 6.556

Data Serving

Massively Parallel Processing Engine [MPP]
Analytical & Modern Data Warehouse [MDW]

Cost for Synapse Analytics Dedicated SQL Pool

- **Region** – EastUS2
- **Tier** – Compute Optimized Gen2
- **DWU** – 300
- **Storage** – 1 TB
- **Monthly Cost** – R$ 14.580

Total Cost for Data Pipelines on Microsoft Azure

- **Storage Layer** = R$ 2.414
- **Data Processing Layer** = R$ 6.556
- **Data Serving Layer** = R$ 14.580

- **Total Monthly Cost** – R$ 23.550

Amazon AWS Big Data Landscape for Data Pipelines

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 & MWAA

Monitoring
Amazon CloudWatch

Cost of a Data Pipeline on Amazon AWS

Data Lake

Raw Repository
Low-Cost Storage

Cost for Amazon S3

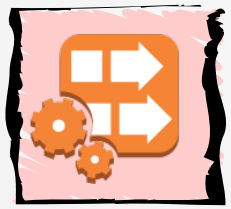
- **Region** – USEast
- **Capacity** – 1 TB
- **Scanned & Returned** – 100 GB
- **Monthly Cost** – R$ 170

Real-Time Ingestion

Ingesting Events at Scale
Store and Replicate Data Securely

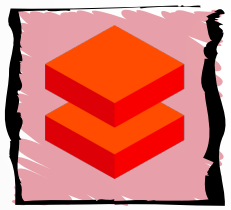
Cost for Amazon Kinesis Firehose

- **Region** – USEast
- **Units** – Thousands
- **Records Amount** – 1 Million
- **Record Size** – 10 KB
- **Conversion** – Parquet & ORC
- **Monthly Cost** – R$ 2.464



Data Processing

Data Analytics Platform
Apache Spark Engine



Cost for AWS Databricks

- **Region** – USEast
- **Workload** – ALL-Purpose Compute
- **Tier** – Enterprise
- **Instance** – M4.XLarge
- **Spec** – 4 vCPUs & 16 GB of RAM
- **VMs** – 3
- **Monthly Cost** – R$ 5.791

Data Serving

Massively Parallel Processing Engine [MPP]
Analytical & Modern Data Warehouse [MDW]

Cost for Amazon Redshift

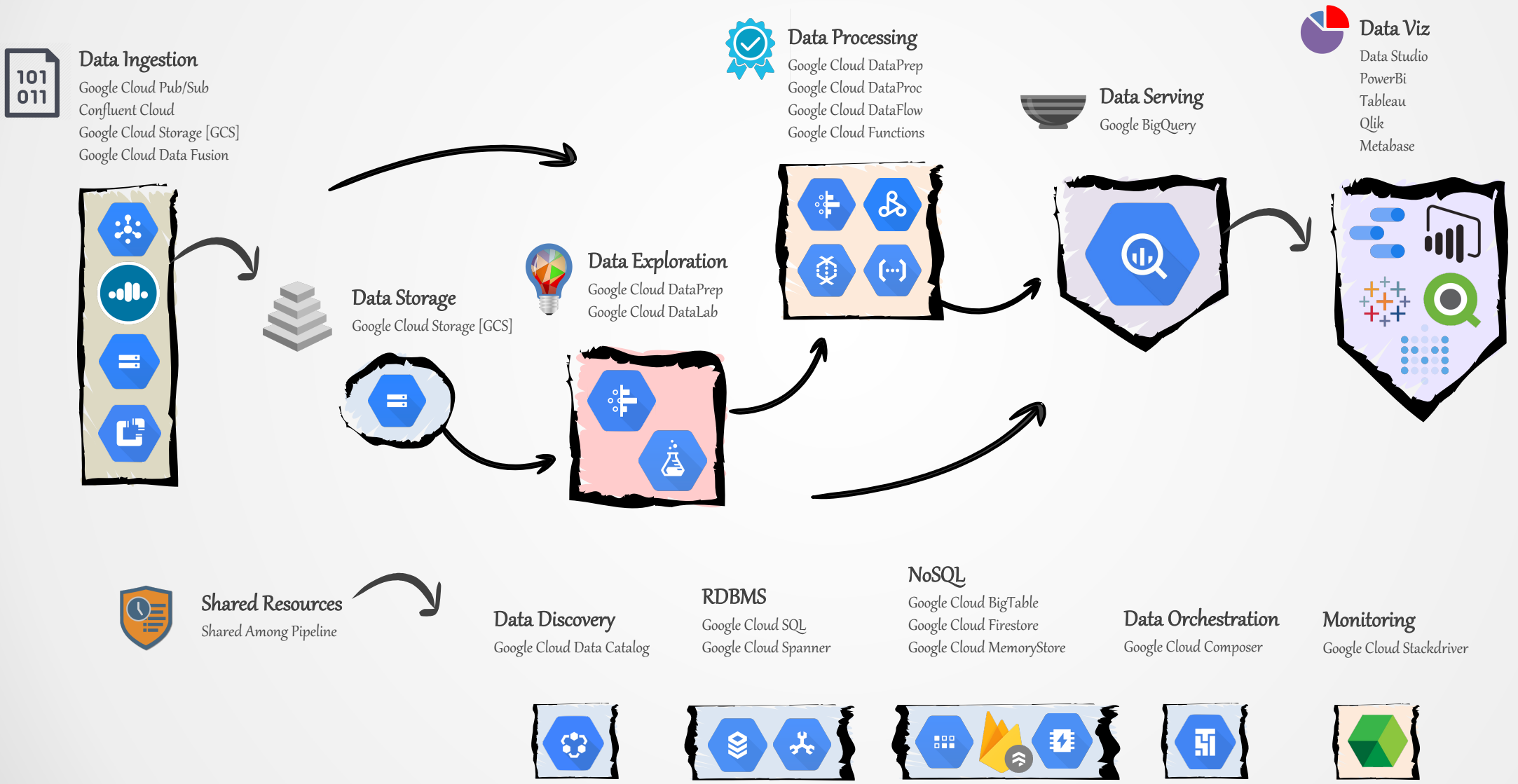
- **Region** – USEast
- **Nodes** – 3
- **Instance** – RA3.XLPlus
- **Spec** – 4 vCPUs & 32 GiB
- **Backup** – 100 GB
- **Spectrum** – 100 GB
- **Storage** – 1 TB
- **Monthly Cost** – R$ 13.238

Total Cost for Data Pipelines on Amazon AWS

- **Storage Layer** = R$ 2.634
- **Data Processing Layer** = R$ 5.791
- **Data Serving Layer** = R$ 13.238

- **Total Monthly Cost** – R$ 21.663

Google GCP Big Data Landscape for Data Pipelines



Cost of a Data Pipeline on Google GCP

Data Lake

Raw Repository
Low-Cost Storage

Cost for Google GCS

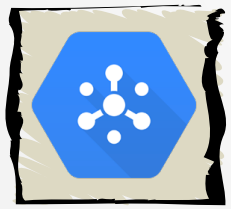
- **Region** – US-Central
- **Capacity** – 1 TB
- **Class A & B Operations** – 1 Million
- **Monthly Cost** – R$ 143

Real-Time Ingestion

Ingesting Events at Scale
Store and Replicate Data Securely

Cost for Cloud Pub/Sub

- **Region** – US-Central
- **Message Volume** – 100 GiB
- **Retained** – 10 GiB
- **Snapshot** – 10 GiB
- **Monthly Cost** – R$ 193



Data Processing

Data Analytics Platform
Apache Spark Engine

Cost for Google DataProc

- **Region** – US-Central
- **VM-Class** – Regular
- **Instance** – N1-Standard-4
- **Spec** – 4 vCPUs & 15 GB of RAM
- **VMs** – 3
- **Storage** – 500 GB SSD
- **Monthly Cost** – R$ 2.156

Data Serving

Massively Parallel Processing Engine [MPP]
Analytical & Modern Data Warehouse [MDW]

Cost for Google BigQuery

- **Region** – US-Central
- **Storage** – 1 TB
- **Streaming Inserts** – 100 GB
- **Queries** – 2 TB
- **Monthly Cost** – R$ 170

Total Cost for Data Pipelines on Google GCP

- **Storage Layer** = R$ 336
- **Data Processing Layer** = R$ 2.156
- **Data Serving Layer** = R$ 170

- **Total Monthly Cost** – R$ 2.662

Cost for a Data Pipeline on Cloud Computing Vendors

Total Cost for Data Pipelines on Microsoft Azure

- **Storage Layer** = R$ 2.414
- **Data Processing Layer** = R$ 6.556
- **Data Serving Layer** = R$ 14.580

- **Total Monthly Cost** – R$ 23,550

Total Cost for Data Pipelines on Amazon AWS

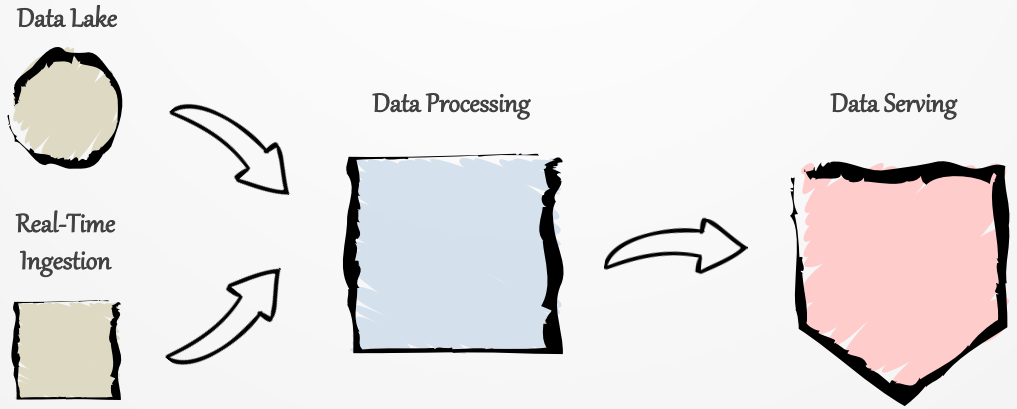
- **Storage Layer** = R$ 2.634
- **Data Processing Layer** = R$ 5.791
- **Data Serving Layer** = R$ 13.238

- **Total Monthly Cost** – R$ 21.663

Total Cost for Data Pipelines on Google GCP

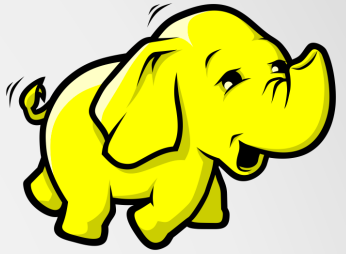
- **Storage Layer** = R$ 336
- **Data Processing Layer** = R$ 2.156
- **Data Serving Layer** = R$ 170

- **Total Monthly Cost** – R$ 2.662




A goal without a plan is just a wish.
Antoine de Saint-Exupéry
quotefancy

OSS Big Data Products on [Spotlight]



Apache Kafka

80% of ALL Fortune 100 Companies Trust. Ingest and Process Data Effortlessly

Apache Pulsar

Messaging & Streaming Platform. Pulsar Functions, Persistent Storage, Multi-Tenancy with Low-Latency

Apache Spark

PySpark, Spark SQL, Java, Scala, R, .NET. Most Used Big Data Product

Apache Airflow

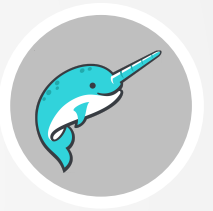
Programmatically Author, Schedule & Monitor Workflows using Python. Newest 2.0 Version Out

Apache Pinot

Real-Time Distributed OLAP Data Store, Designed for Low-Latency Queries at Scale

Trino

Data Processing Engine Unleashing SQL at Scale & Providing Data Virtualization Process Layer



Dremio

Next-Generation Data Lake Engine for Interactive Query in a Blazing Fast Speed

YugaByteDB

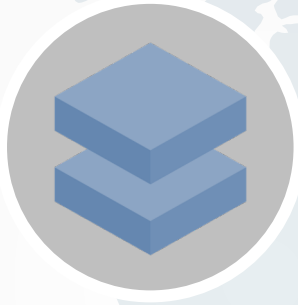
Cloud-Native Database Platform using Different APIS – Redis, Postgres & Cassandra

Azure Big Data Products on [Spotlight]

Azure Purview

Unified Data Governance with Data Discovery, Sensitive Data Classification & End-to-End Data Lineage

1. Data Discovery, Classification and Mapping
2. Data Catalog: Searching & Web-Based Experience
3. Data Governance: Enabling Key Insights and Understanding of Data Quality Rules



Azure Databricks

Fast, Easy & Collaborative Apache Spark Based Analytics Service Providing Fast Deployment Process

1. Databricks Runtime ~ Optimized for Cloud Storage
2. Managed Delta Lake
3. Integrated Workspace – GitHub
4. Production Jobs & Workflows
5. Enterprise Security
6. Integrations using ODBC & JDBC
7. SQL Analytics – Redash + Delta Lake Engine

Azure Synapse Analytics

MDW with Limitless Analytics Service with Unmatched Time to Insight – PaaS & SaaS Approaches

1. Serverless & Dedicated Options
2. Data Lake Exploration
3. Code-Free ETL & ELT
4. Deeply Integration with Apache Spark & SQL Engines
5. Languages – T-SQL, Python, Scala, Spark SQL and .NET
6. Cloud-Native HTAP with Azure Synapse Link ~ CosmosDB
7. AI & BI

AWS Big Data Products on [Spotlight]

Managed Streaming for Apache Kafka [MSK]

Fully Managed Service to Build and Run Applications ~ Apache Kafka to Process Streaming Data Effortlessly

1. Amazon MSK Runs and Manages Apache Kafka, Maintain Open-Source Compatibility, MirrorMaker, Apache Flink, and Prometheus
2. VPC Network Isolation, AWS IAM for Control-Plane API Authorization, Encryption at Rest, TLS Encryption & In-Transit

Amazon Glue & DataBrew

Serverless Data Integration Service for ETL, ELT, Catalog, Lineage & Transformations for Cleaning and Enriching Data

1. Discover, Prepare, & Combine Data for Analytics, Machine Learning, and Application Development
2. DataBrew is New Visual Data Preparation Tool ~ Clean and Normalize Data for Analytics and Machine Learning

Amazon Managed Workflows for Apache Airflow [MWAA]

Managed Orchestration Service for Apache Airflow, Operate End-to-End Data Pipelines in Cloud at Scale with Minimum Effort & Configuration

1. Data Secured by Default Running in an Isolated and Secure Cloud Environment using VPC, Data is Automatically Encrypted using KMS
2. Connect ~ AWS or On-Premises Resources Required for Workflows Including Athena, Batch, Cloudwatch, DynamoDB, DataSync, EMR, Fargate, EKS, Firehose, Glue, Lambda, Redshift, SQS, SNS, Sagemaker & S3

GCP Big Data Products on [Spotlight]

Cloud Data Fusion

Fully Managed, Cloud-Native Data Integration at Any Scale using Ephemeral DataProc Cluster Underneath

1. Code-Free ETL & ELT Deployment of Data Pipelines
2. Library of 150+ Configured Connectors & Transformations
3. Built with OSS Core CDAP for Pipeline Portability

Cloud Dataflow

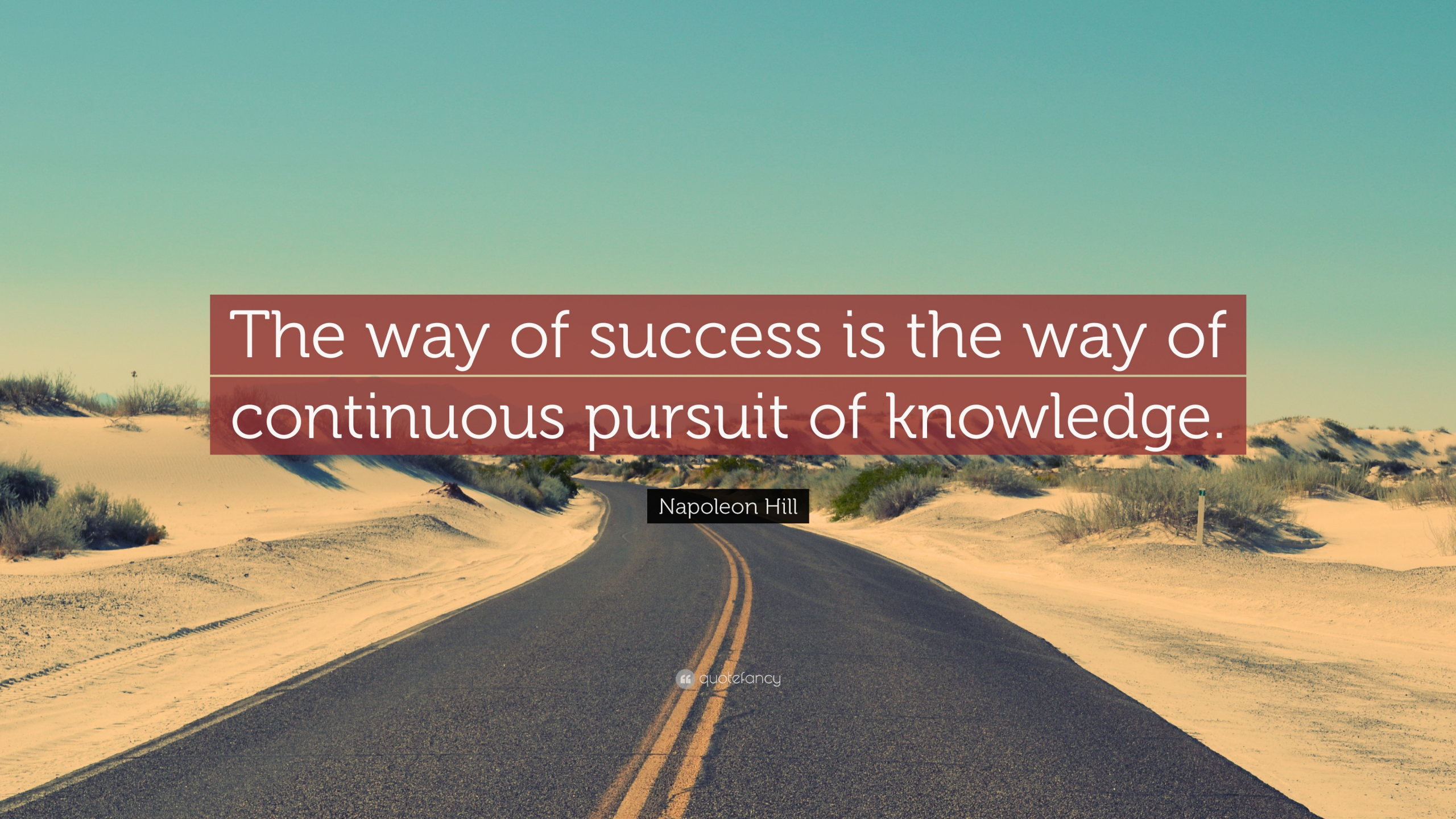
Unified Stream and Batch Data Processing Serverless, Fast, and Cost-Effective using Beam Framework

1. Automated Provisioning and Management of Processing Resources
2. Horizontal Autoscaling of Worker Resources
3. OSS Community-Driven Innovation with Apache Beam SDK
4. Reliable and Consistent Exactly-Once Processing

BigQuery

Serverless, Highly Scalable, and Cost-Effective Multi-Cloud Data Warehouse Designed for Business Agility

1. Analyze Petabytes of Data Using ANSI SQL at Blazing-Fast Speeds, with Zero Operational Overhead
2. Democratize Insights with a Trusted and Secure Platform Scales
3. Gain Insights from Data Across Clouds with a Flexible, Multi-Cloud Analytics Solution ~ Omni


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.”

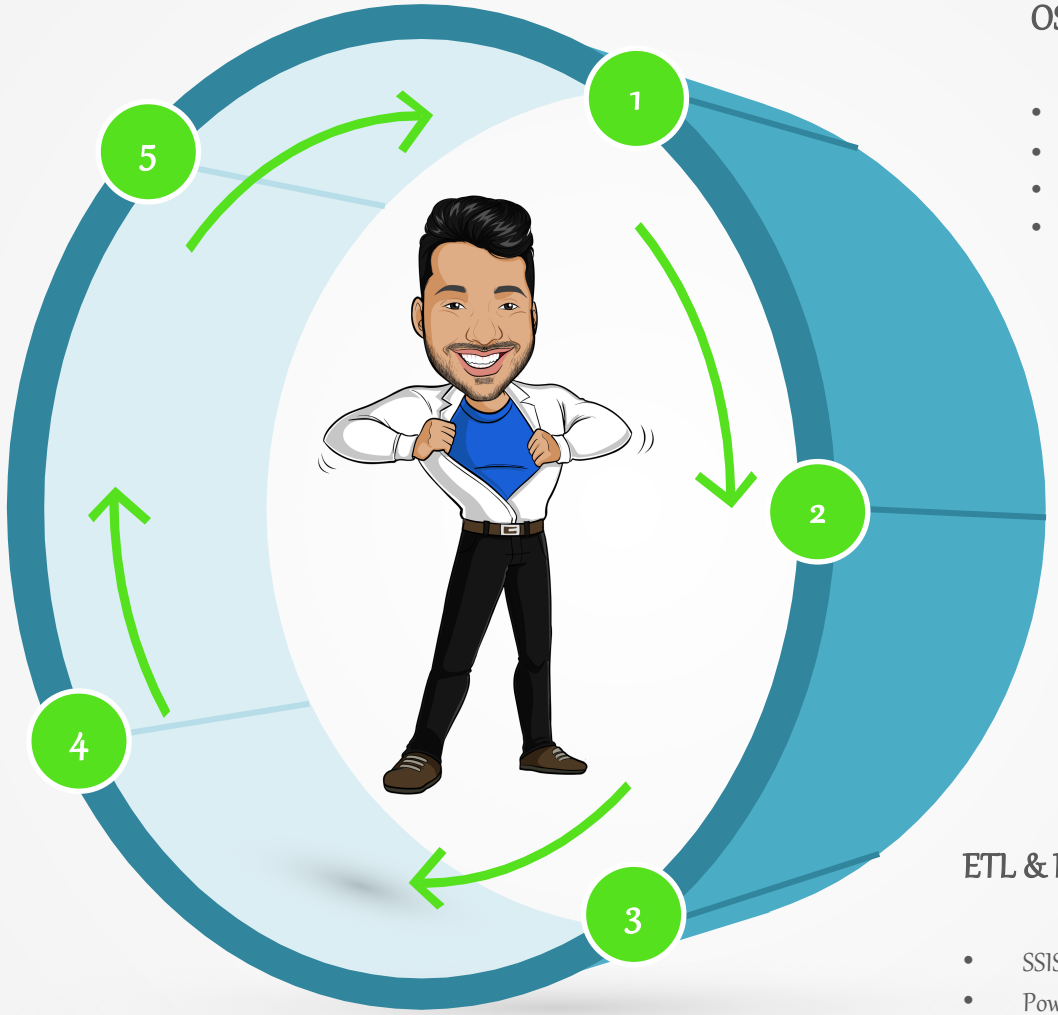
- Gordon Lindsay Glegg

“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

DBMS & NoSQL

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

ETL & DW

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

Data Pipelines & Cloud Computing

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

Distributed Systems & Big Data Frameworks

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

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 analysis 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

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

Python for Everybody Specialization

introduce fundamental programming concepts including data structures, networked application program interfaces, and databases, using the python programming language.

Data Engineer

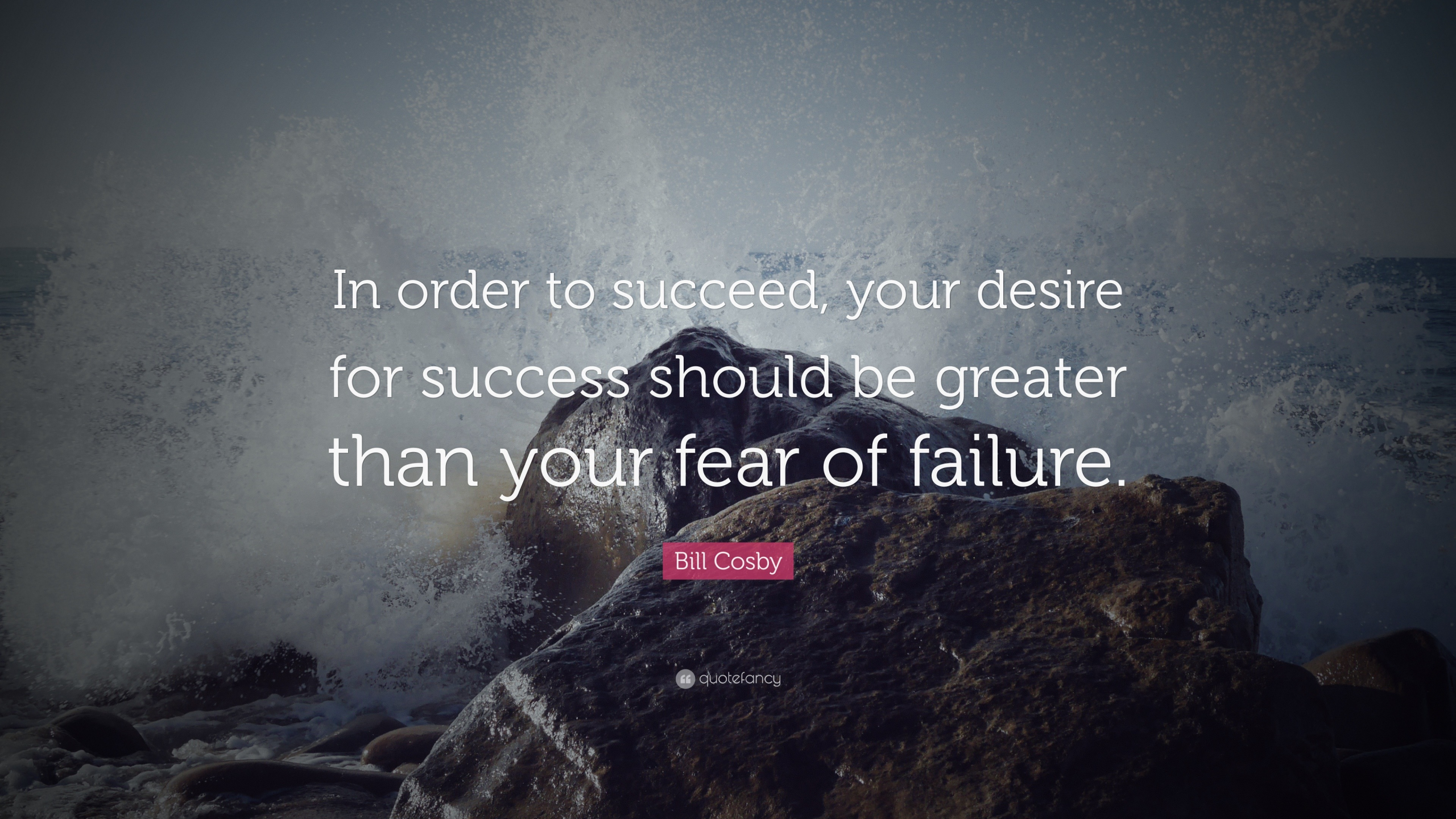
learn to design data models, build data warehouses and data lakes, automate data pipelines, and work with massive datasets.

Architecting with GKE Specialization

how to implement solutions using gke, including building, scheduling, load balancing, and monitoring workloads, as well as providing for discovery of services, managing role-based access control and security, and providing persistent storage to these applications.

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

Podcast
engenhariadedadoscast

Thank You

One Way Solution

ONEWAY

SOLUTION