Modern Data Pipelines for Analytics

HELLO!

- Data Consultant - Data Engineer at Servian, Sydney
- Google Certified Professional Data Engineer
- Master of Data Science, Monash University, Australia
- Bachelor of Computer Science, R.V.C.E Bangalore, India

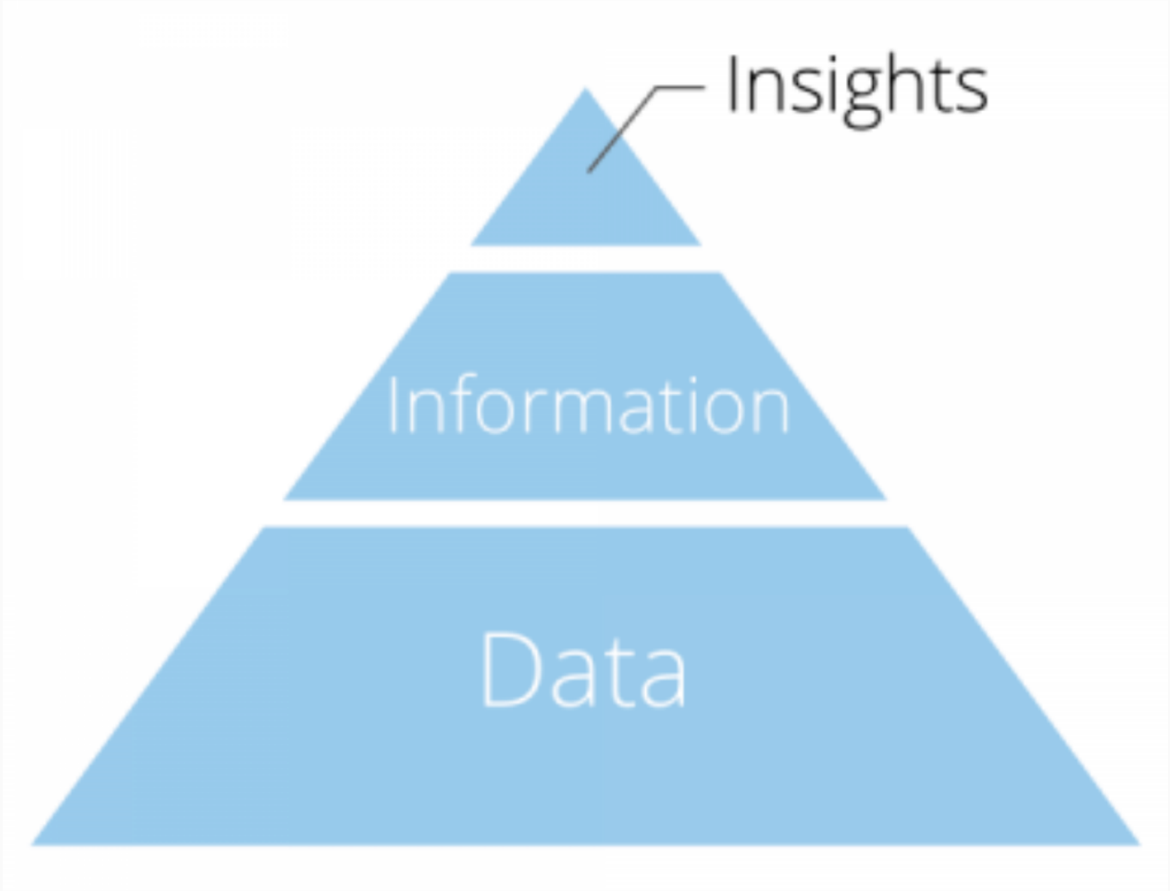
Mayana Mohsin Khan

www.linkedin.com/in/mayanamohsinkhan

Whats for the day?

- Data & Data Pipeline
- Data Governance from Data Engineering prespective
- ETL vs ELT
- Traditional vs Modern Data Pipeline
- Modern Data Pipelines on AWS/GCP/Azure
- Case Study: Building ML Pipeline with GCP.

Introduction

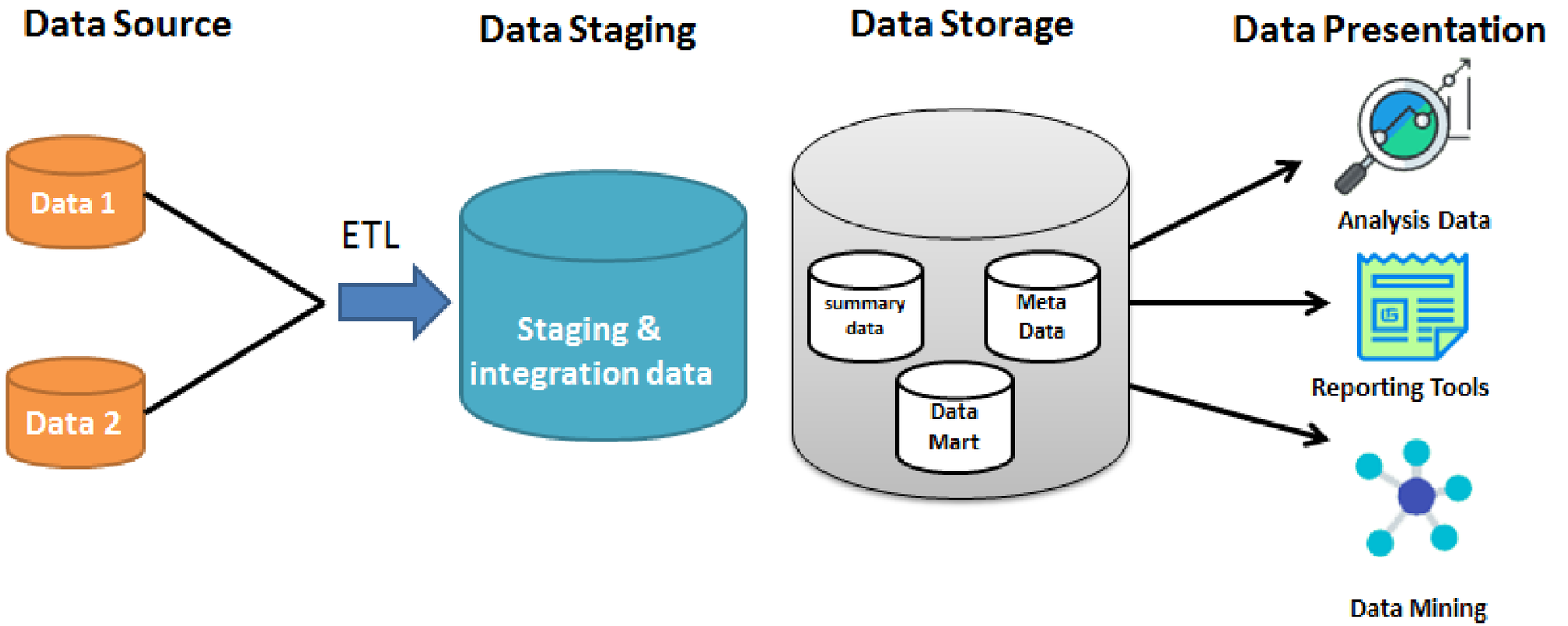


Data Pipeline an Introduction

Operations:

1. Data Cleaning
2. Data Governance
3. Data Enrichment
4. Data Processing

Data Pipeline an Introduction



Data Governance

Data Governance

- Improves understanding of data.
- Improves data quality.
- Data transparency.
- Increase use of data to make program and policy decisions.
- Enable easier access to data.
- Enable consistent and high-quality data releases.
- Increase data security.
- Reduce operational friction.
- Protect the needs of data stakeholders.
- Reduce costs

Data Governance

Data Management Body of Knowledge (DMBoK)

ETL
(Extract Transform and Load)

1. Data is transformed at staging area
2. Required a specialised in transformation engine.
3. Strengths: Facilitates data quality, data security and data compilance

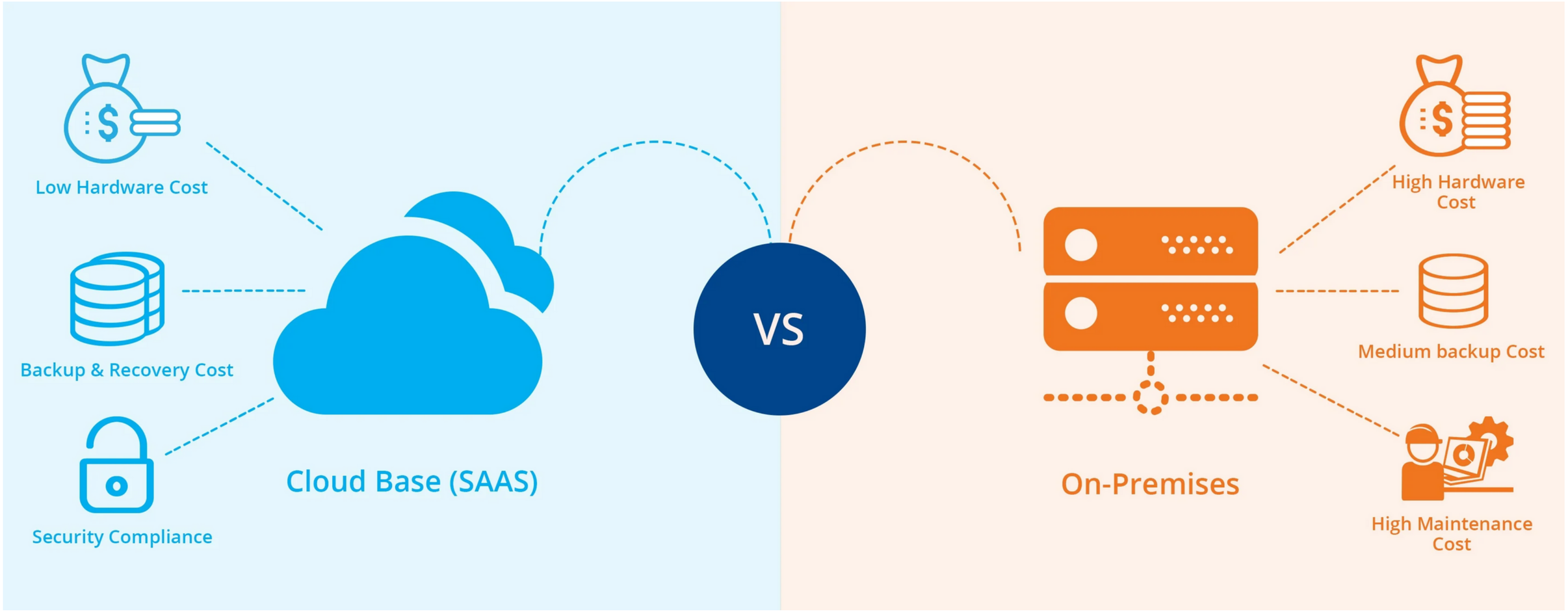
ELT
(Extract Load and Transform)

1. Data is transformed within the data warehouse.
2. Resources native to data warehouse are used for computation and processing.
3. Strengths: More Flexible, greater availability, Scalability,

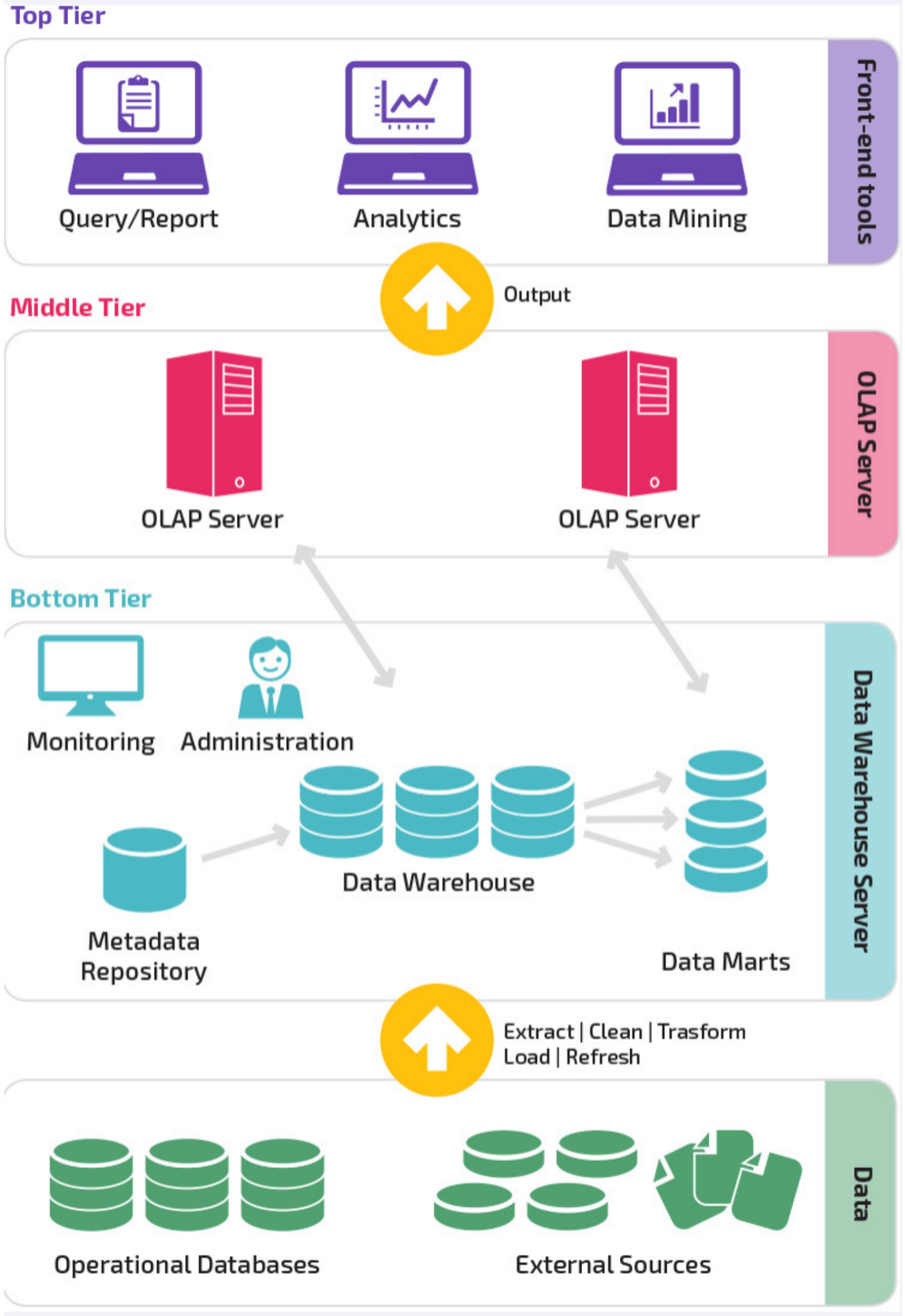
ETL vs ELT

- More flexibility
- Greater accessibility
- Scalability
- Faster load times
- Faster transformation times
- Less time required for data maintenance

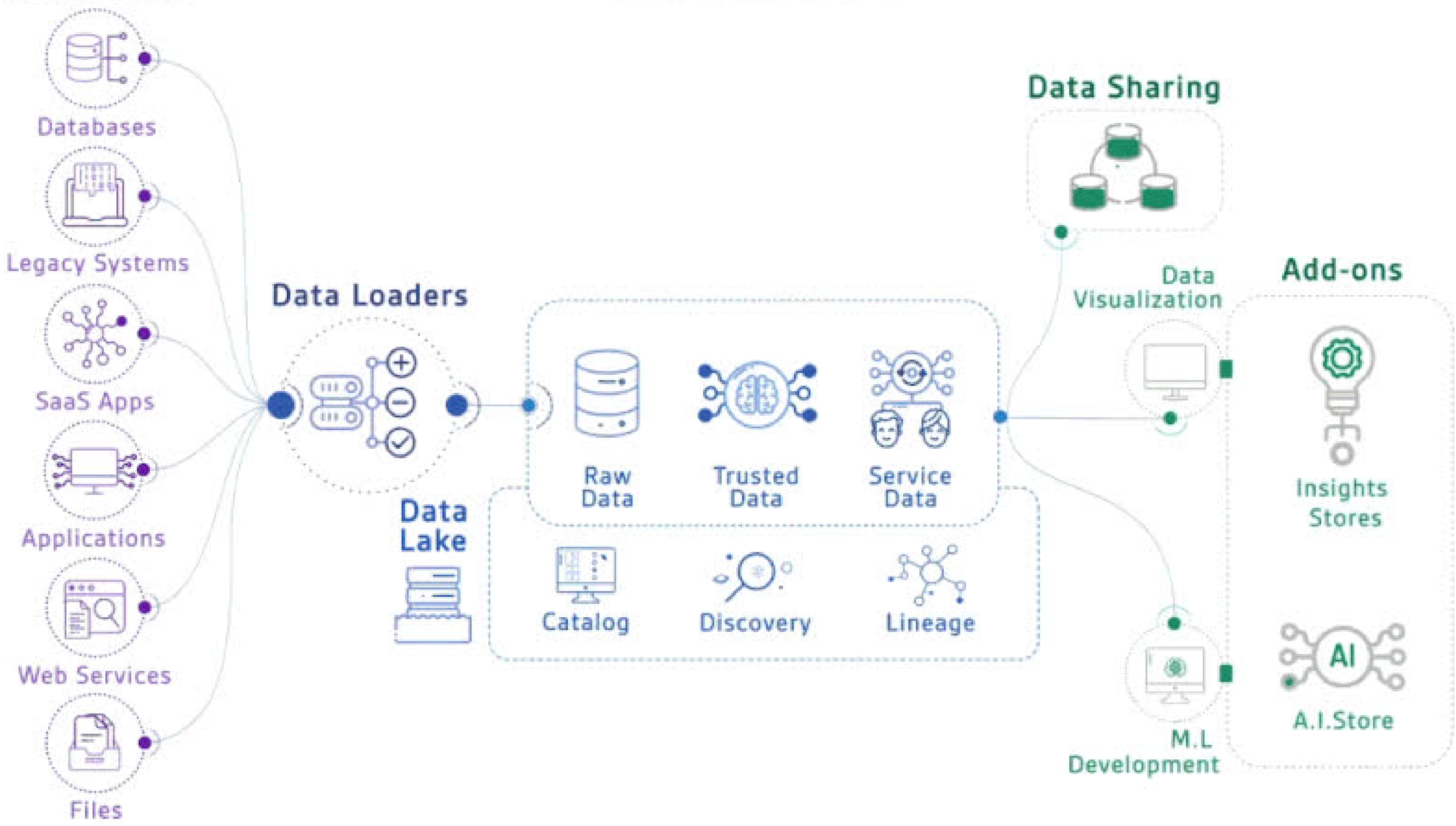
Cloud Based VS On-Premise



Traditional Data Pipelines



Modern Data Pipelines



Features of Modern Data Pipelines

- SaaS Offerings
- Scalable Architecture
- Fault-Tolerant & disaster recovery
- Data Encryption by default
- Pay as you use billings
- Easily process large Volumes of Data
- Real-time/Batch Data Processing and Analytics
- Streamlined Data Pipeline Development

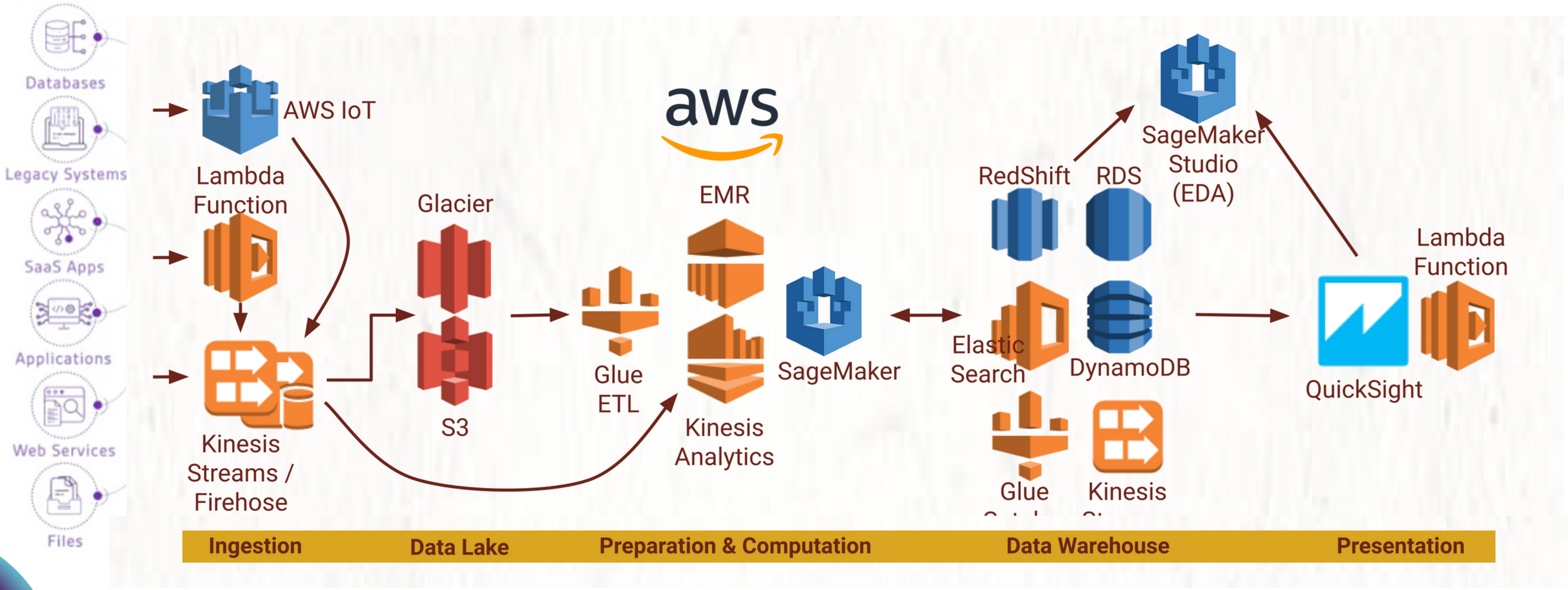
snowflake®

amazon REDSHIFT

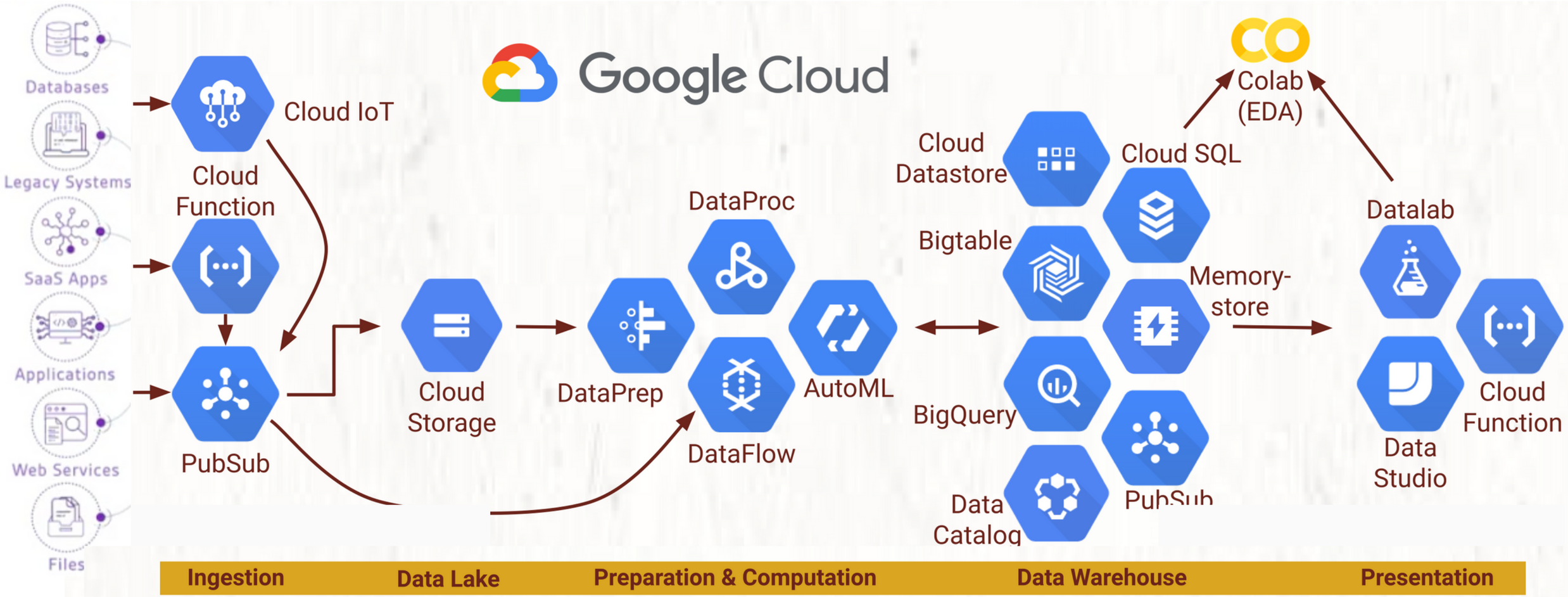
FIREBOLT

BigQuery

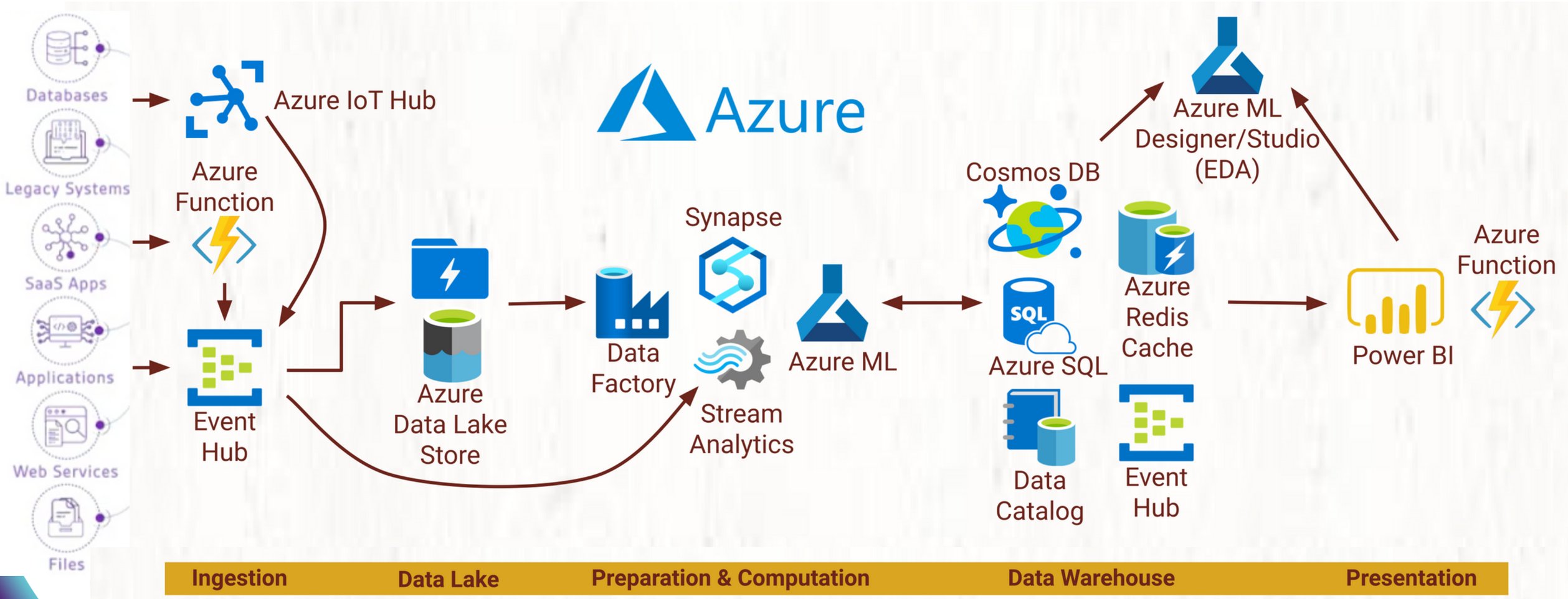
Building Modern Data Pipelines with AWS



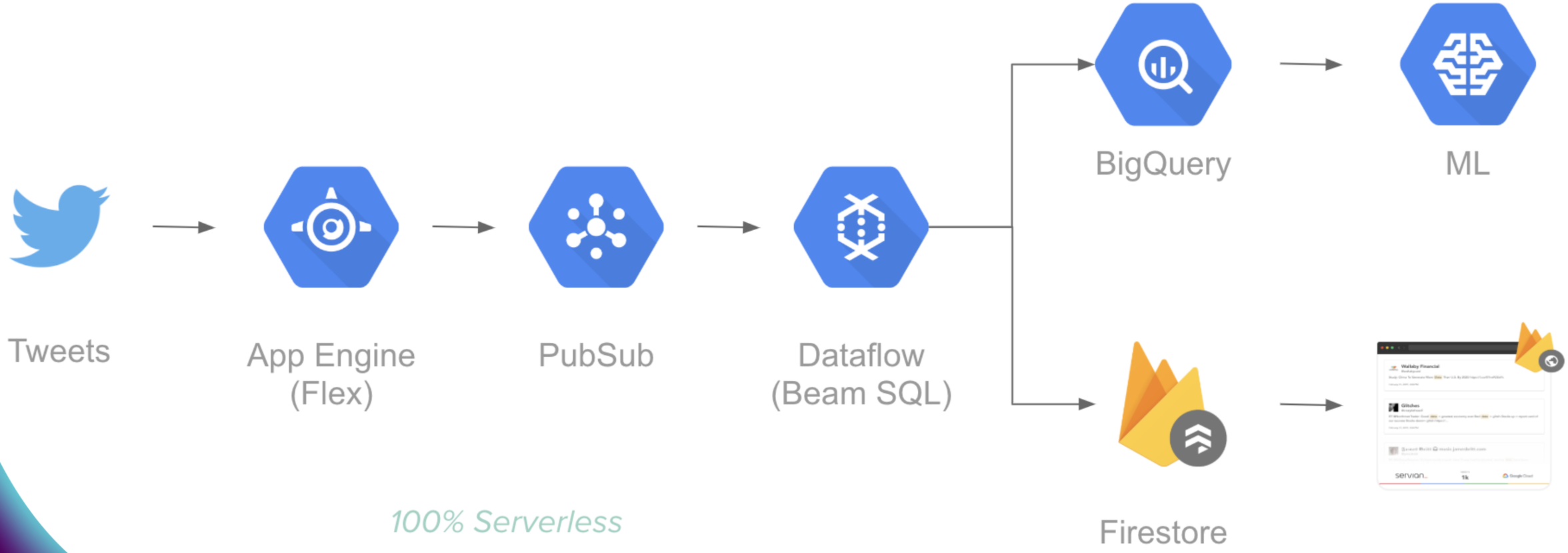
Building Modern Data Pipelines with GCP



Building Modern Data Pipelines with Azure



Case Study: Building ML Pipeline with GCP.



100% Serverless

Thank You