Mahidol University
Wisdom of the Land

BIG DATA PROCESSING

Week 7 Pipeline

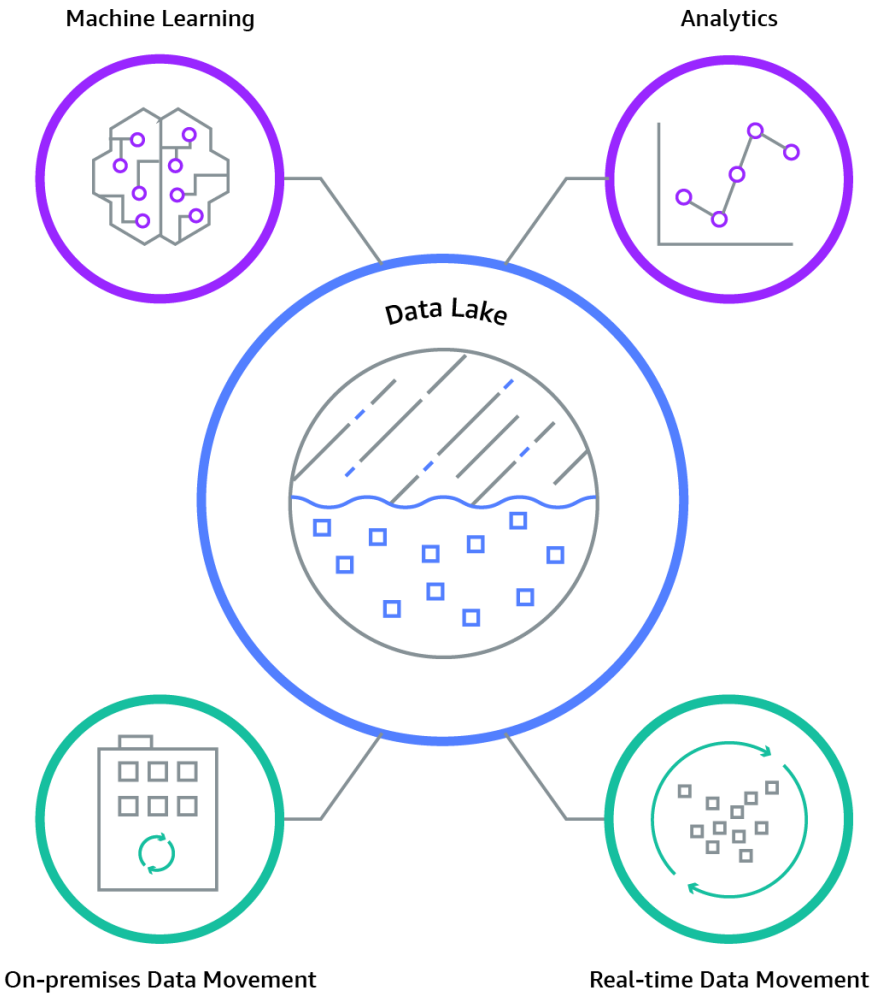
Mahidol University
Wisdom of the Land

Hadoop as a Big Data Platform

- Hadoop is one of **the most favourite**
  - Storage (HDFS)
  - Processing/Analysis (Hive, Spark, etc.)
  - Ecosystems (Sqoop, KafKa, AirFlow)
- Good choice for a data lake
  - Cheap storage compares with others
- On premise **=> need large investment**
- **Need installation/administration**
- Con=> tie storage & processing together

Mahidol University
Wisdom of the Land

Data Lake



Mahidol University
Wisdom of the Land

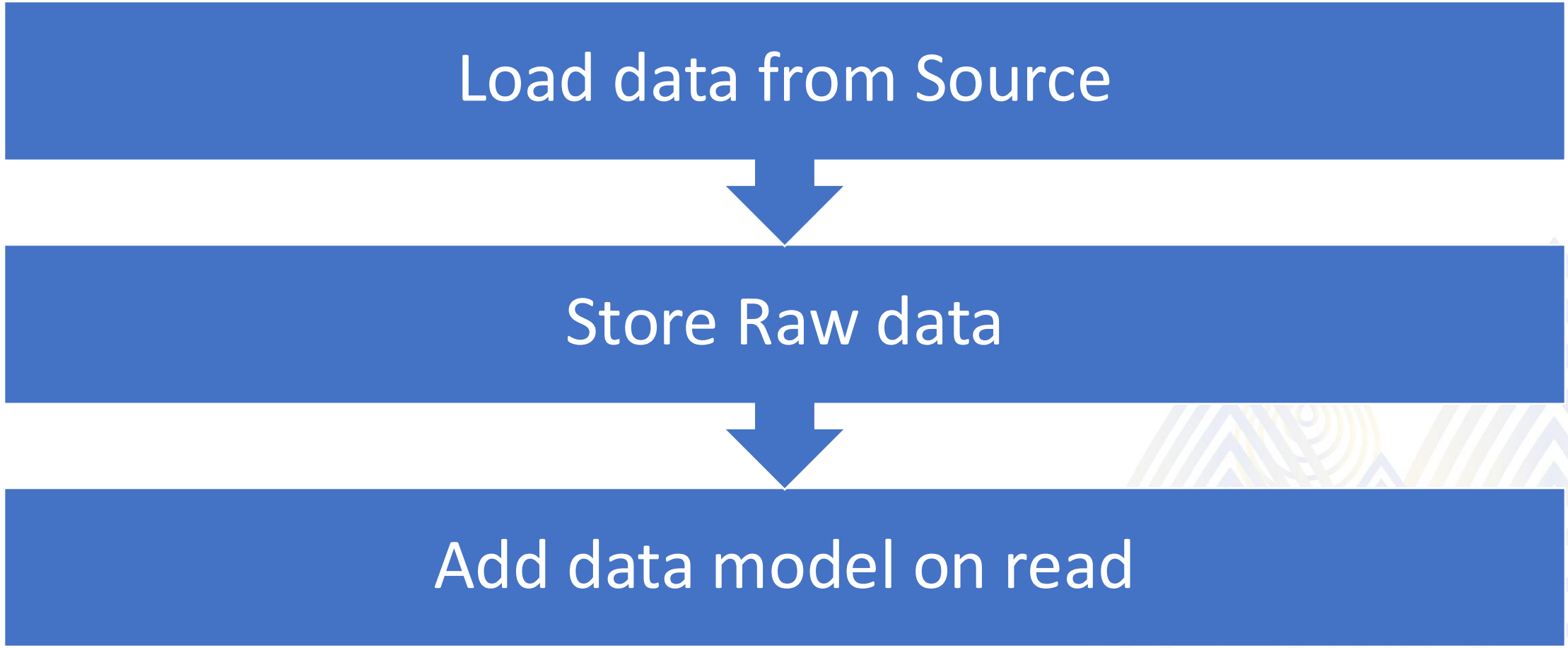
Data Lakes

- Infrastructure that many stream can flow into
- Stored for processing in the original form
- Massive storage with huge processing power

> “If you think of a data mart as a store of bottled water, cleansed and packaged and structured for easy consumption,
> the data lake is a large body of water in a more natural state.
> The contents of the data lake stream in from a source to fill the lake,
> and various users of the lake can come to examine it, dive in, or take samples. ”-- James Dixon(2010), the Pentaho Corporation's CTO

Mahidol University
Wisdom of the Land

Data Lakes Process



Schema on read

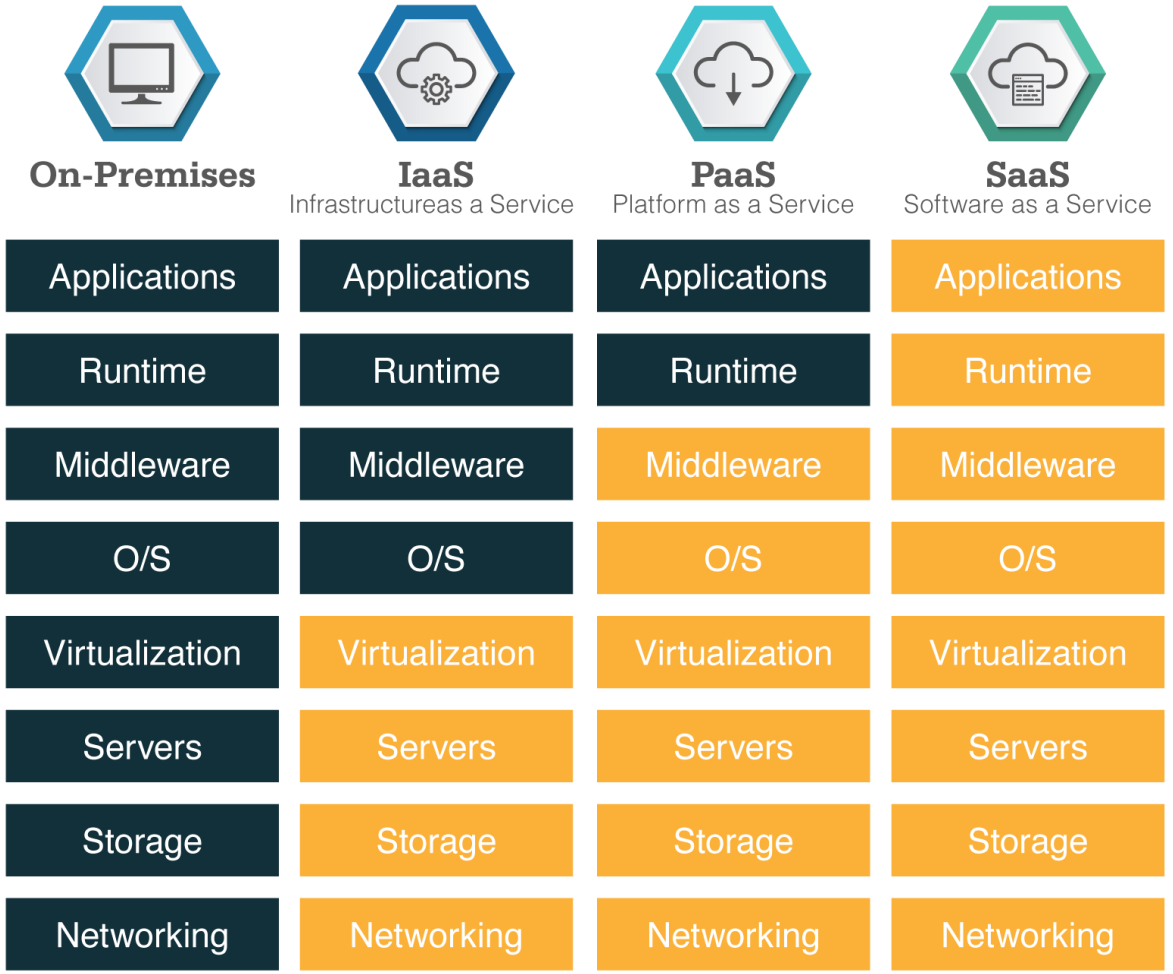
Mahidol University
Wisdom of the Land

Use cloud storage as a data lake

- Cheap storage for data archive
- Cost per GB/month
- High availability
- High durability
- Focus on warm/cold data
- Transaction data
- Encrypt data for security

Mahidol University
Wisdom of the Land

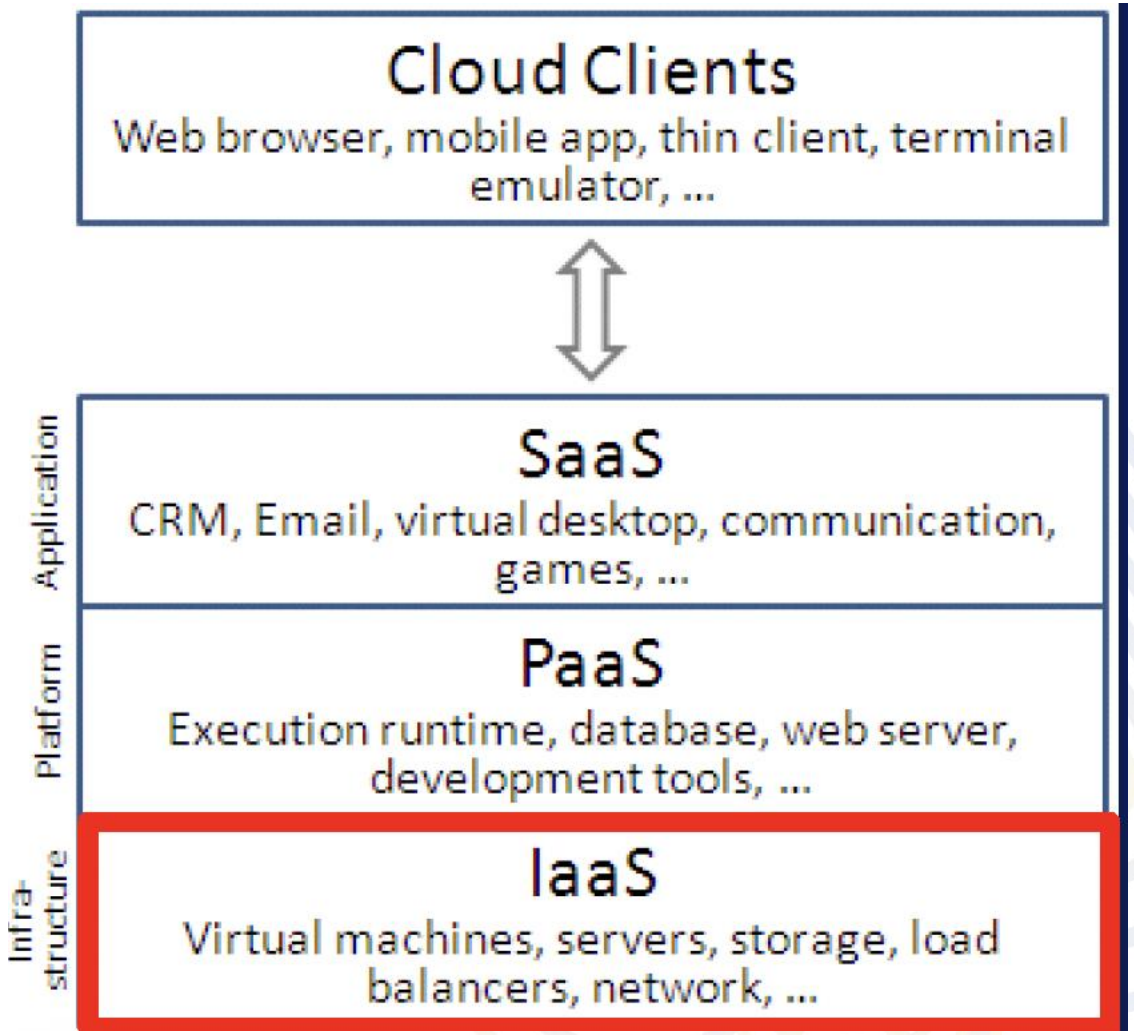
Cloud Service Model



Mahidol University
Wisdom of the Land

IaaS

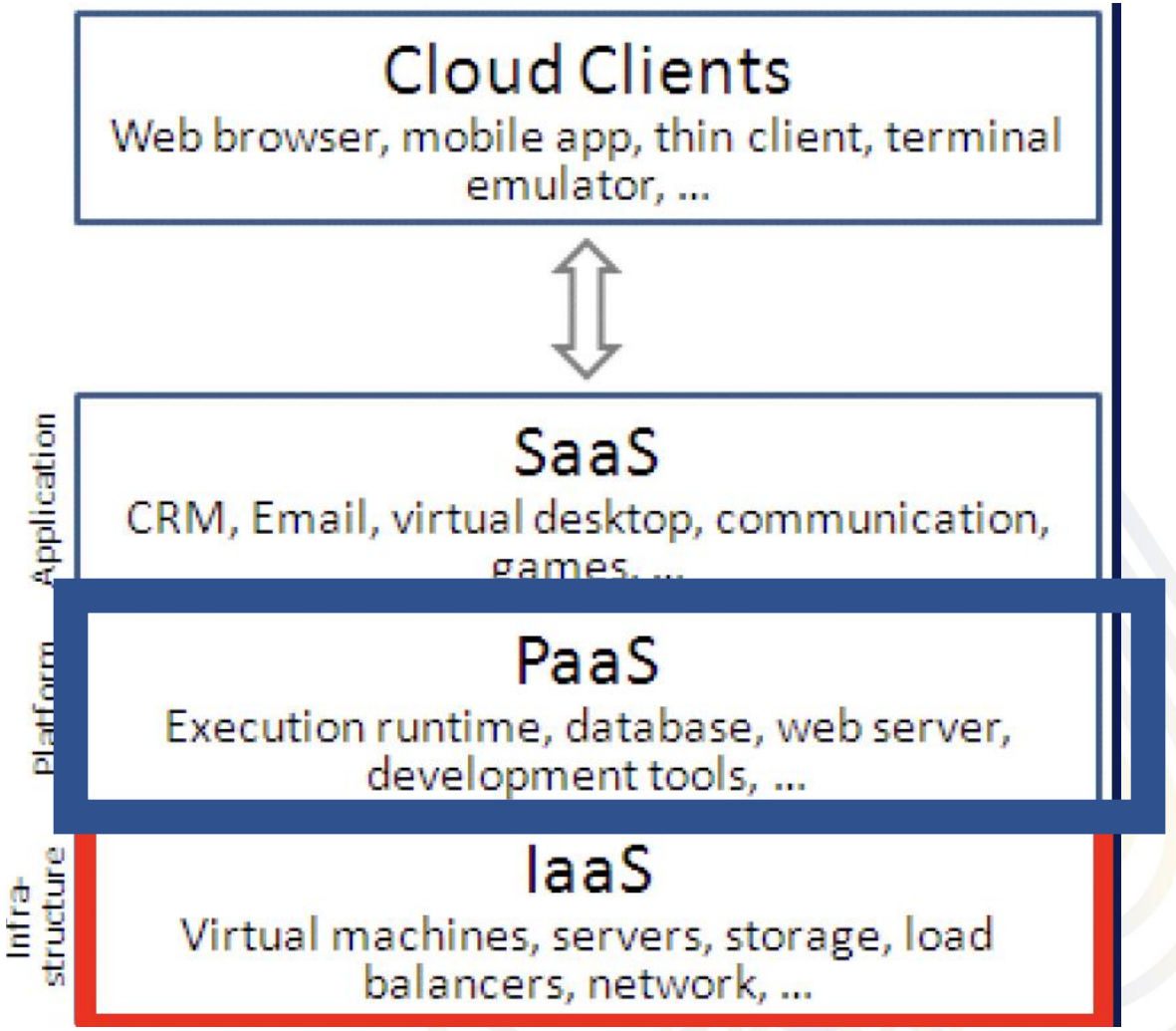
- Infarstructure as a Service
- Install & maintain OS by yourself
  - Amazon EC2
  - Virtual Machine
  - Server
  - Load balancer
  - Network



Mahidol University
Wisdom of the Land

PaaS

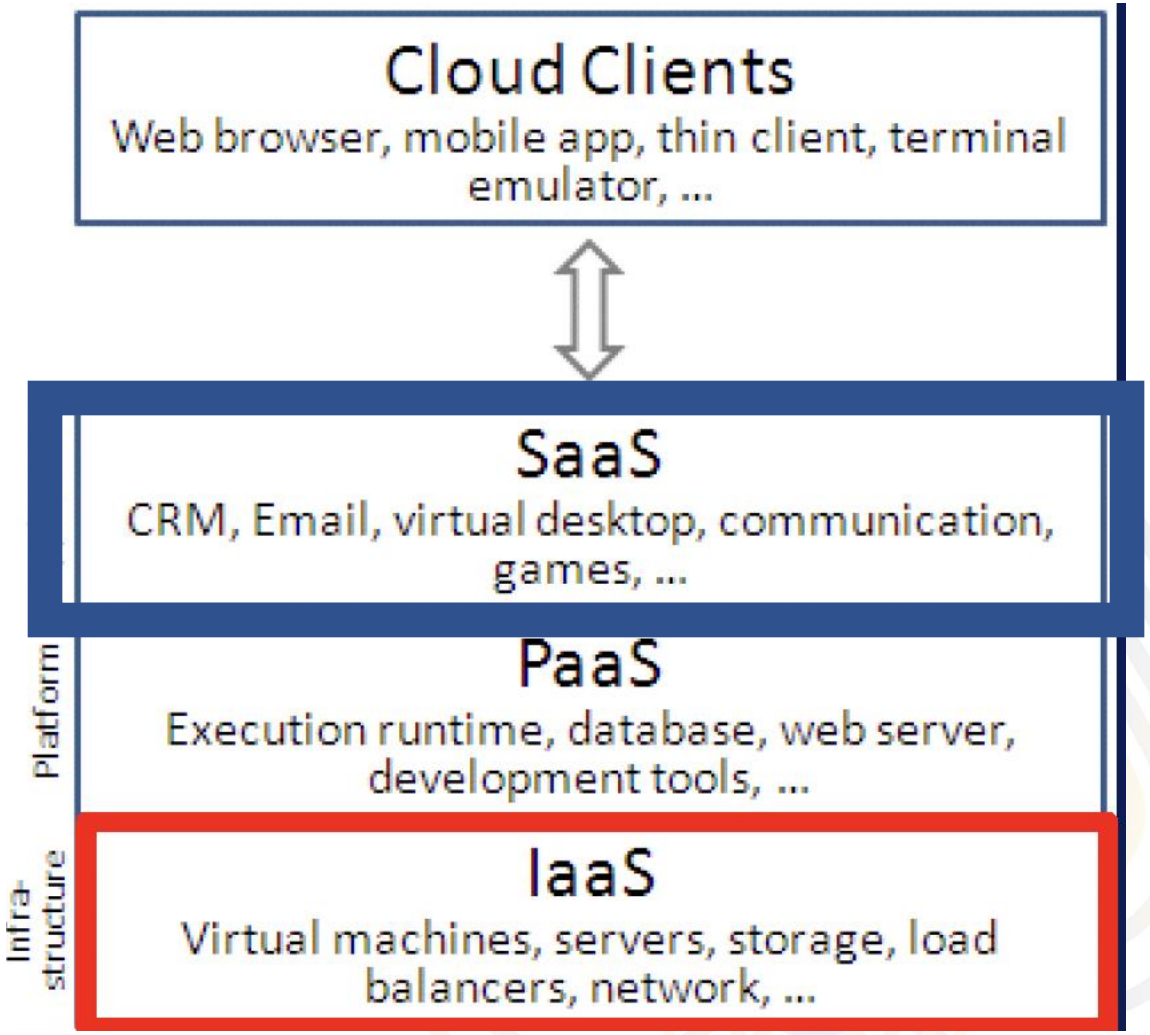
- Platform as a Service
- Install Application by yourself
  - Google App, Microsoft Azure
  - Execution runtime
  - Database
  - Web server
  - Development Tools



Mahidol University
Wisdom of the Land

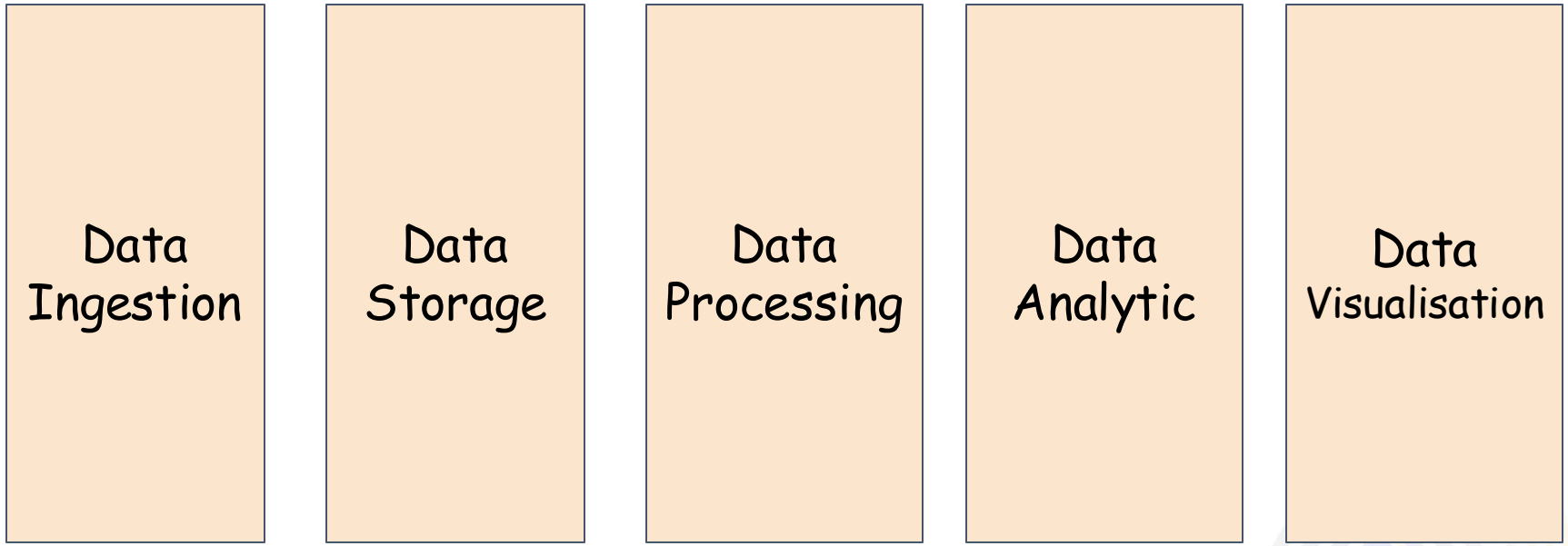
SaaS

- Software as a Service
- Just use an application
  - Dropbox, Google App



Mahidol University
Wisdom of the Land

Typical Big Data Batch Pipeline on Cloud



Ingestion Services On Cloud

(PaaS)

Cloud Storage (IaaS)

RDBMS/Data Warehose as a Service

Hadoop/Spark on Cloud
Presto on Cloud
ML as a Service
(PaaS)

Tableau
Power BI

(SaaS)

Mahidol University
Wisdom of the Land

Spark/Hadoop as a service for analytics

- Separate process layer from storage layer
- No need to install/admin a cluster
- Use processing tools: Hive, Spark, MLlib
- Start cluster only processing time, then terminate cluster when finish.
- Scalable CPU powers
- Pay only processing (CPU) time.

Mahidol University
Wisdom of the Land

Analytics as a service

Amazon EMR

Cloud Dataproc

Big Data

Mahidol University
Wisdom of the Land

Data Warehouse as a Service

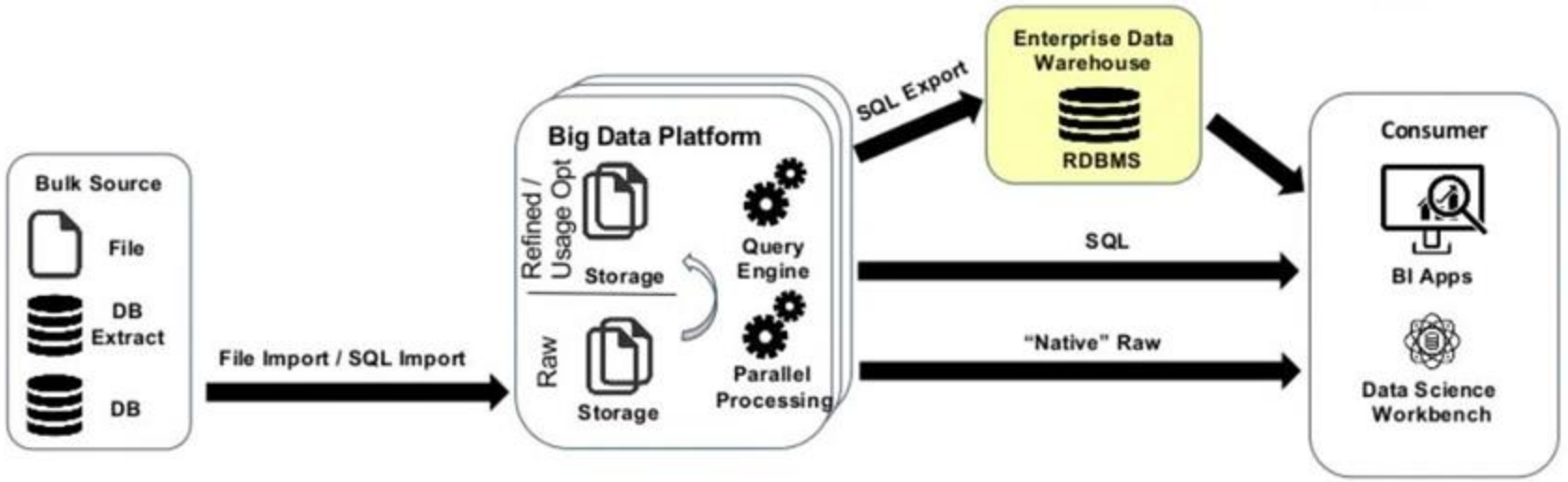
- A fully-managed and cloud-based interactive query service for massive datasets
- Petabyte Data WareHouse
- Example Cloud Services
  - Google BigQuery
  - Amazon Redshift
  - Azure Synapse

Mahidol University
Wisdom of the Land

Fast SQL as a Service

- MPP SQL
- Example Cloud Services
  - Google BigQuery
  - Amazon Athena
  - Azure Data Lake Analytics

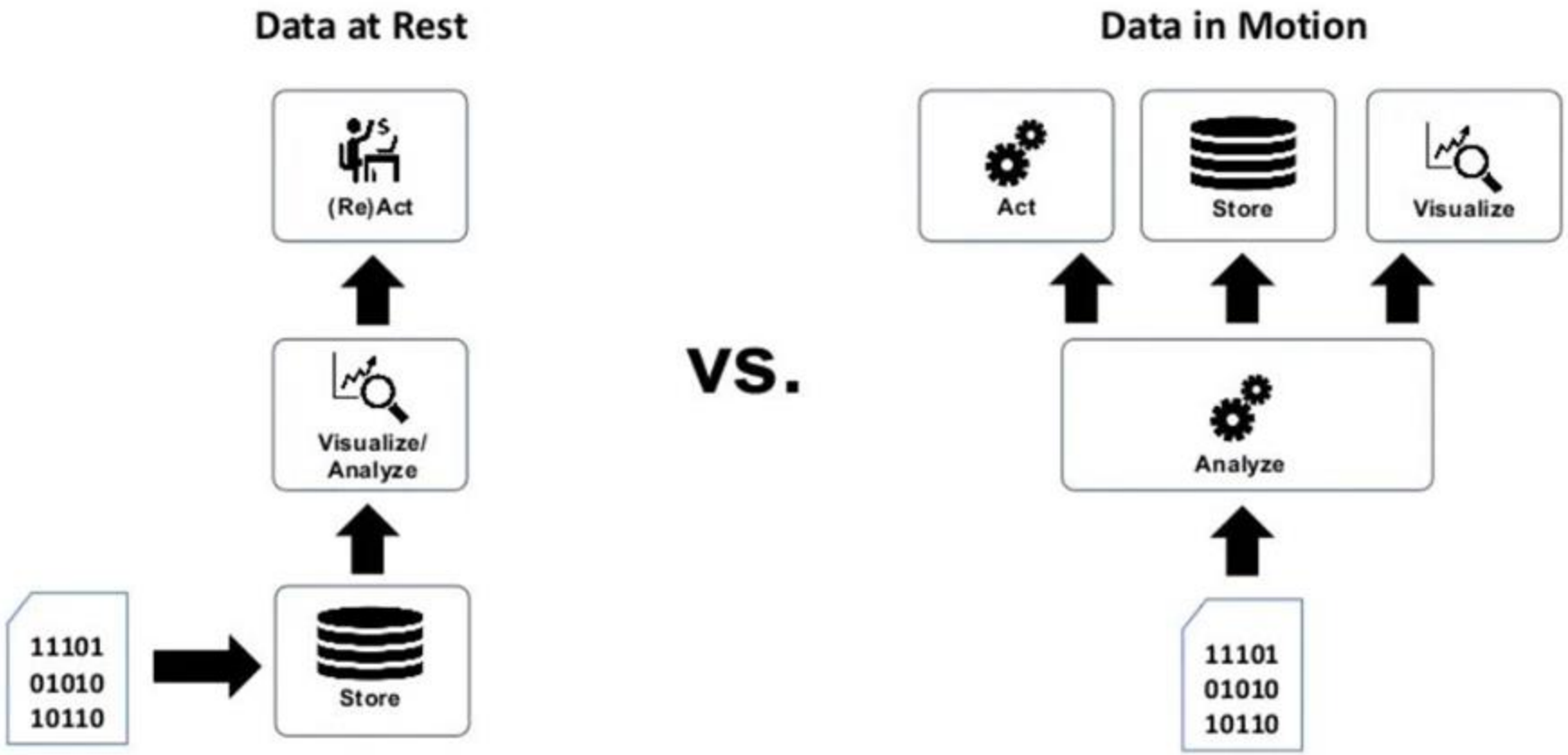
1) Data Lake + DW Architecture



Low latency

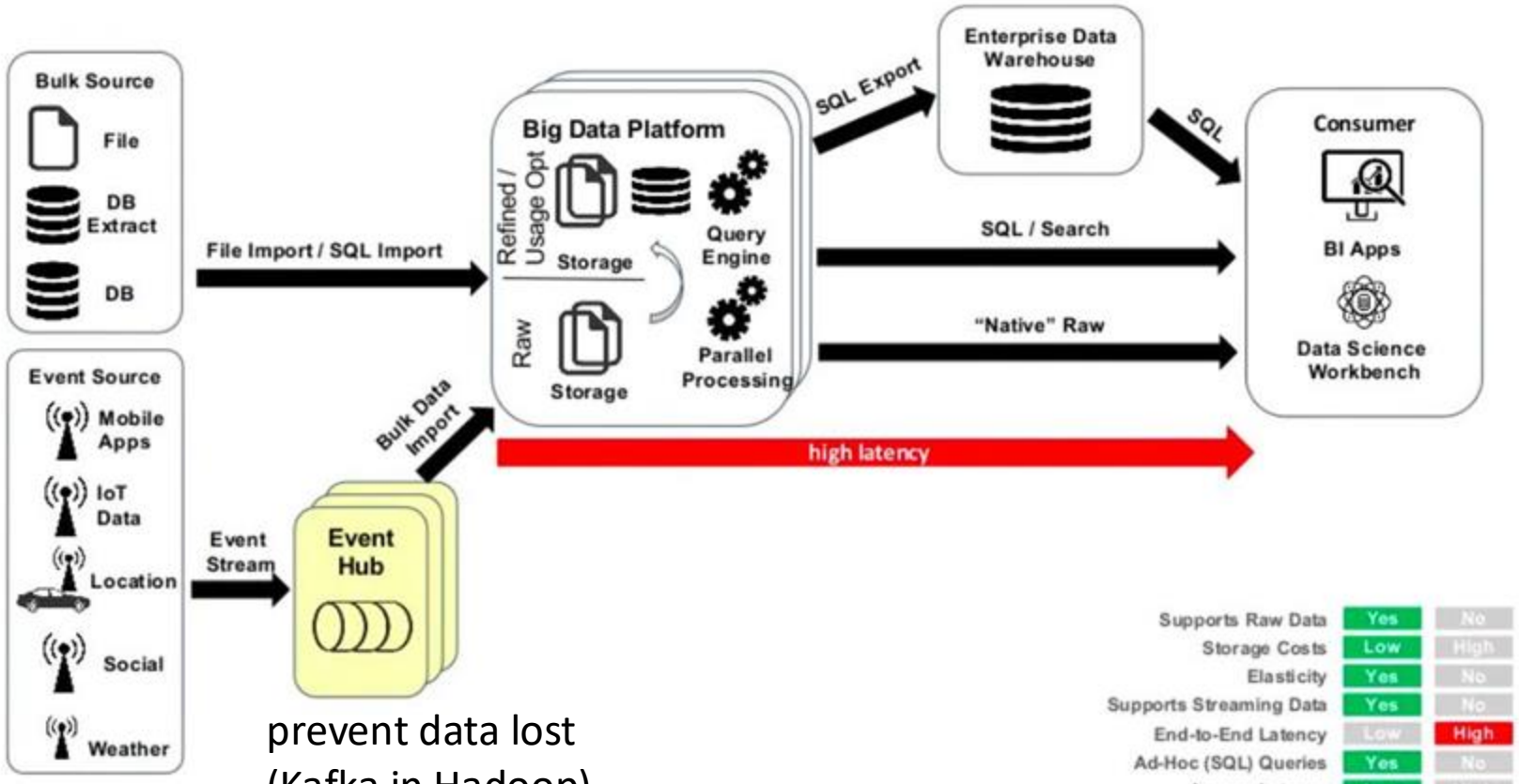
Supports Raw Data	Yes	No
Storage Costs	Low	High
Elasticity	Yes	No
Supports Streaming Data	Yes	No
End-to-End Latency	Low	High
Ad-Hoc (SQL) Queries	Yes	No
Access Latency	Low	High

Source: Fundamentals Big Data and AI Architecture, Guido Schmutz



Source: Fundamentals Big Data and AI Architecture, Guido Schmutz

2) Big Data Architecture with Streaming

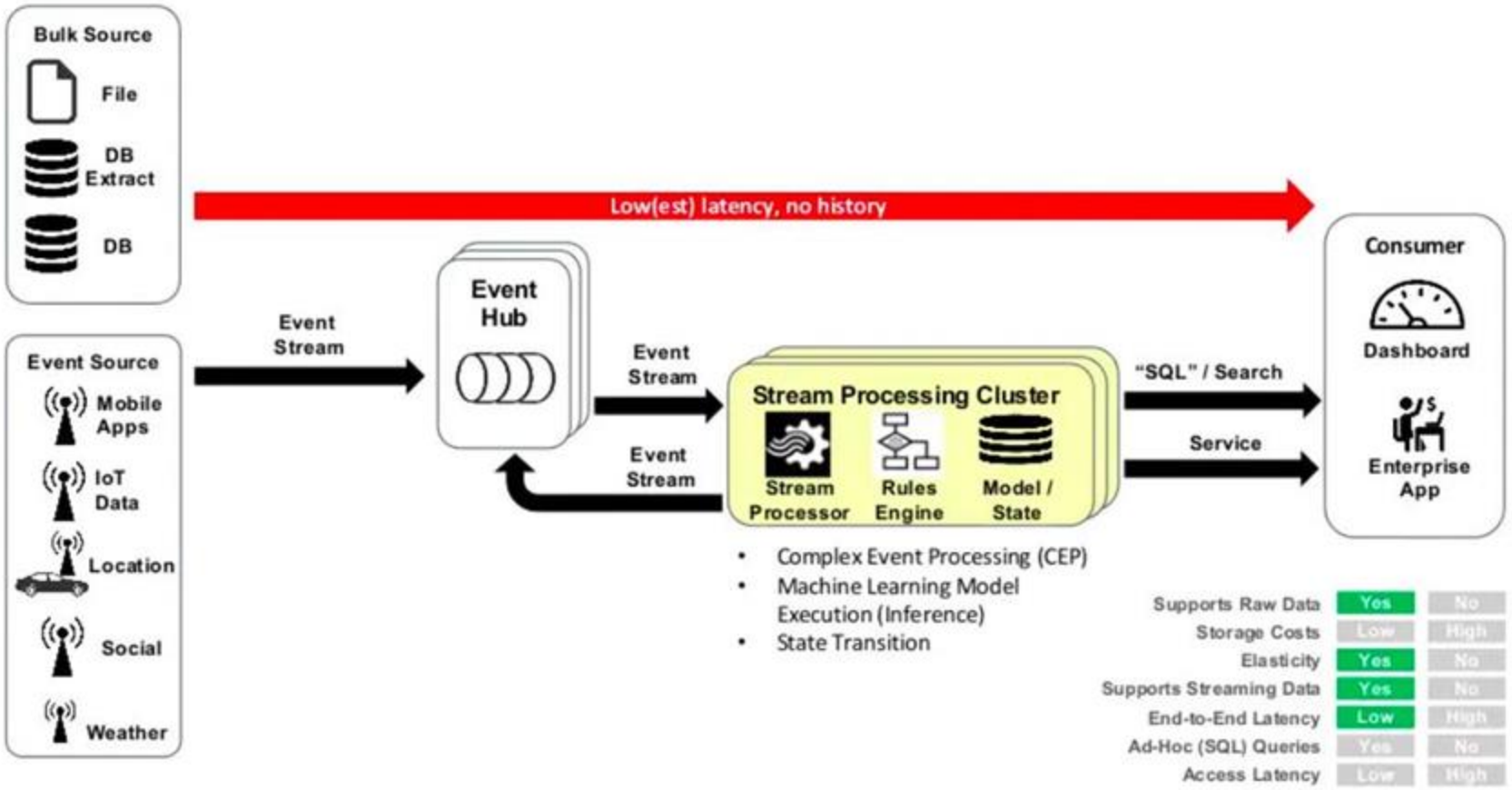


prevent data lost
(Kafka in Hadoop)

Source: Fundamentals Big Data and AI Architecture, Guido Schmutz

Mahidol University
Wisdom of the Land

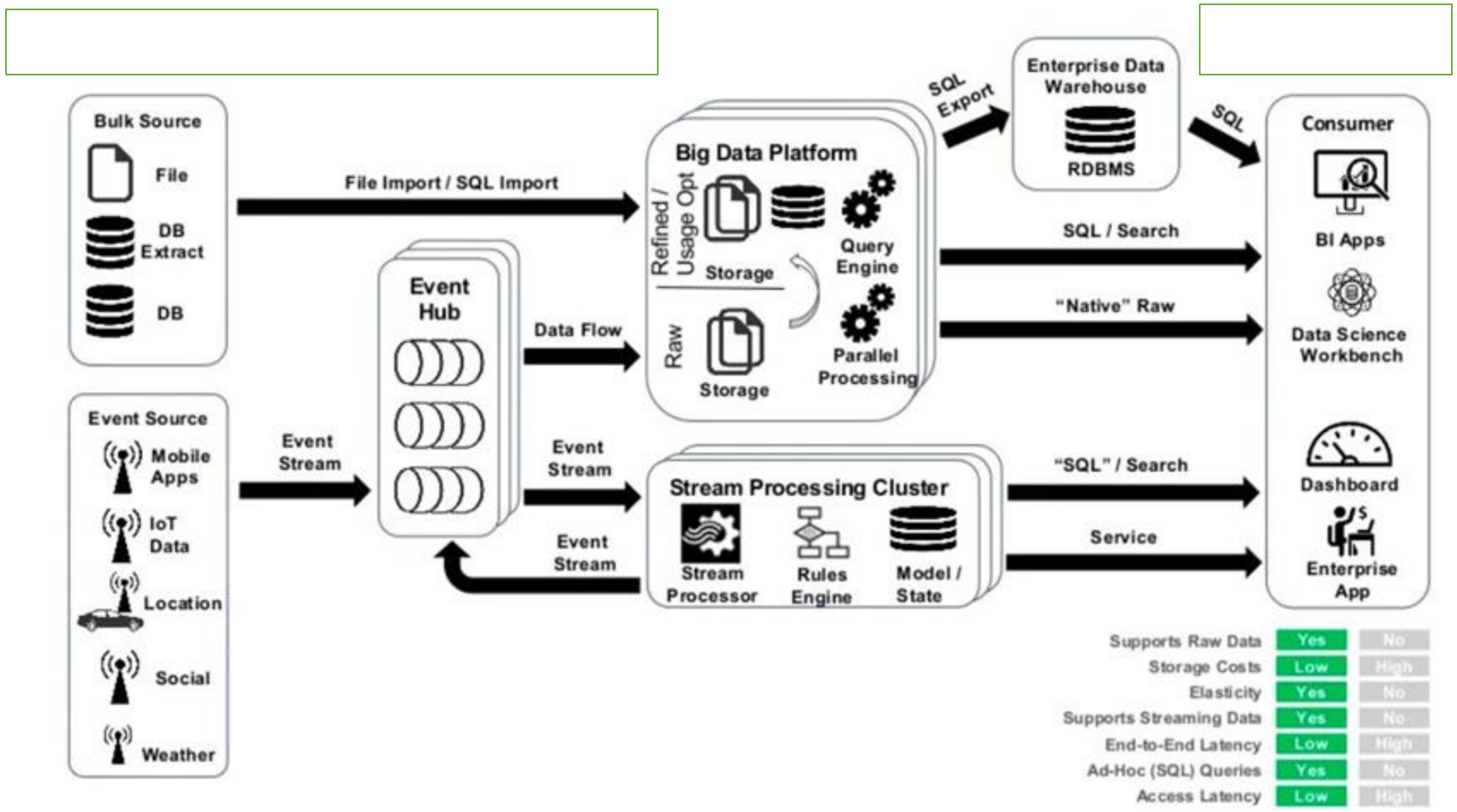
Event Processing Architecture



Source: Fundamentals Big Data and AI Architecture, Guido Schmutz

Mahidol University
Wisdom of the Land

3) Data Lake + Event Processing



Source: https://t.co/pqXT5ZOaad

Data Visualization

Power BI

ORACLE DATA VISUALIZATION

tableau SOFTWARE

Google Data Studio

Data Processing / Data Analytics

Batch

Scriptin

SQL

In-

NoSQL

Streamin

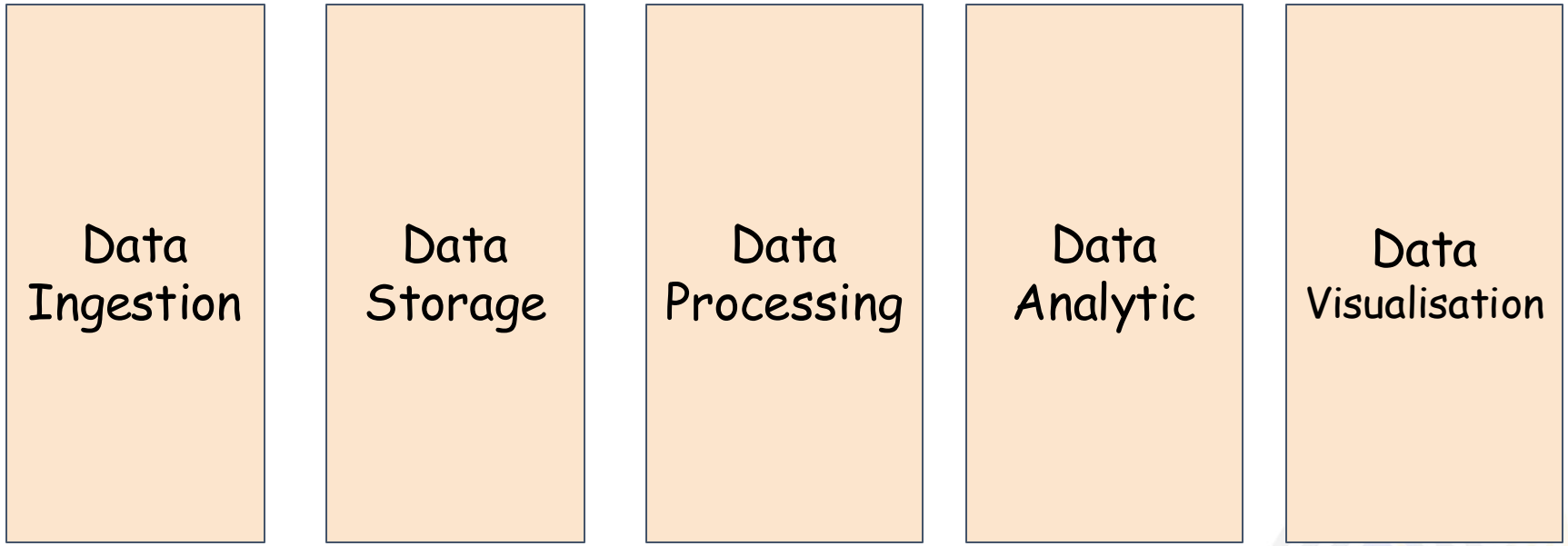
Machine Learning

Data Storage

Data Ingestion

Mahidol University
Wisdom of the Land

Typical Big Data Batch Pipeline



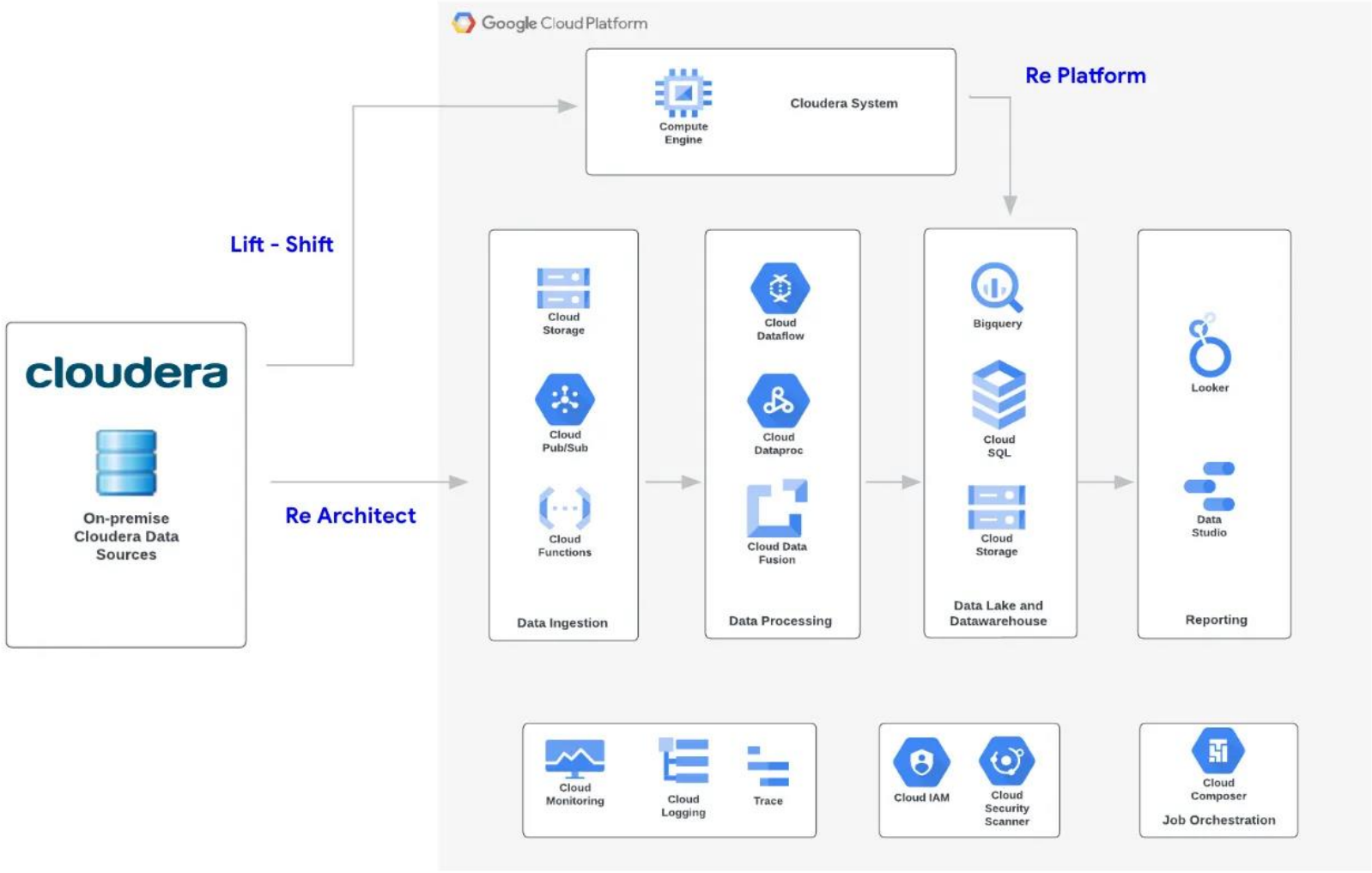
Sqoop
KafKa
Aegisthus

Hadoop HDFS
S3
Cassandra
MySQL Cluster
Elastic search
Redis

Spark
Storm
Spark Streaming
Hive, Pig
Presto

Tableau

On-Premise v.s. Public Cloud Services



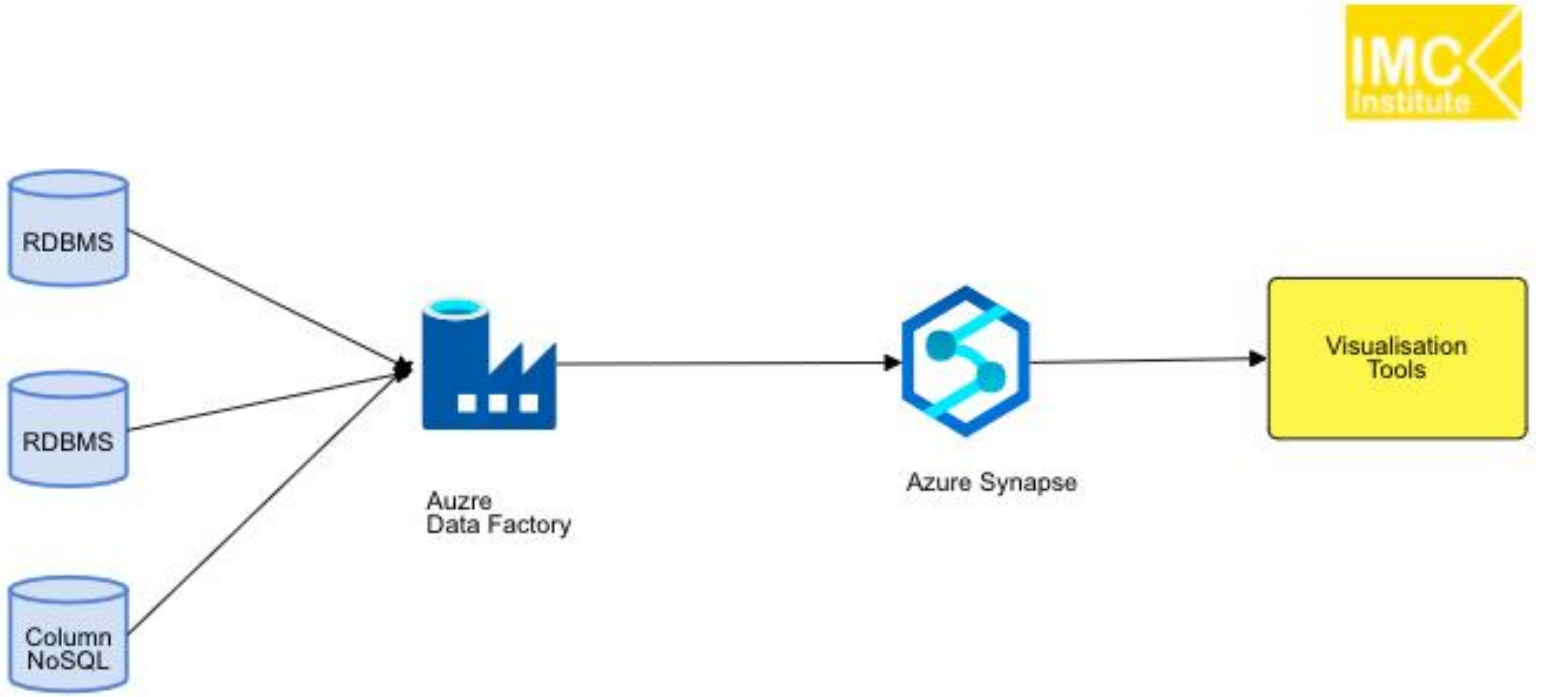
https://blog.searce.com/migrating-hadoop-workloads-to-google-cloud-platform-9fca2e821d07

Mahidol University
Wisdom of the Land

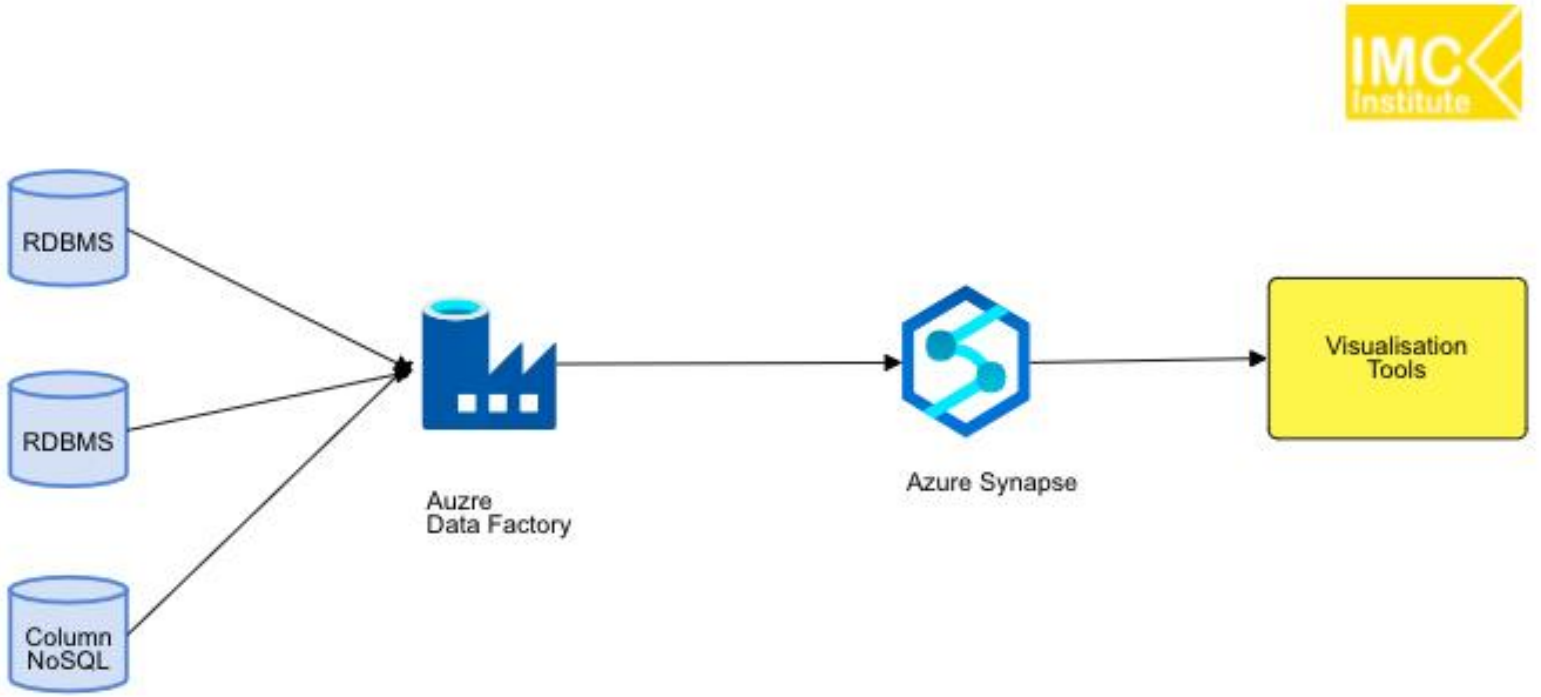
On-Premise v.s. Public Cloud Services

Provided by: IMC Institute		On-Premise	GCP	AWS	Azure	Huawei Cloud	Alibaba Cloud	Oracle Cloud
Ingestion	Event	KafKa	Cloud Sub/Pub	Managed Streaming for KafKa	Apache Kafka on HDInsight	Distributed Message Service (DMS) for Kafka	DataHub	Streaming
		Flume		Kinesis	Event Hub	Data Ingestion Service (DIS)		
	ETL	Sqoop / Nifi	Cloud DataFusion	Glue	Data Factory	Data Replication Service (DRS)	Data Integration	Data Integration
Storage	Data Warehouse	Data Warehouse	Google BigQuery	Redshfit	Synapse	Data Warehouse Service (DWS)	AnalyticDB	Autonomous Data Warehouse
	Data Lake	Hadoop HDFS	Cloud Storage	S3	Azure Data Lake	Object Storage Service (OBS)	Object Storage Service	Object Storage
	Cold Data Archive		Cold Line Cloud Storage/ Arcive storage	S3 Glacier	Archive Storage			Archive Storage
	Hot data	RDBMS	Cloud SQL	RDS	Azure Database	RDS	ApsaraDB RDS	Autonomous Database
		NoSQL	Cloud Datastore/ Cloud BigTable	DynamoDB	Cosmos DB	Gauss DB	Tablestore	NoSQL
Processing / Analytics	Generic	Hadoop/Spark	Cloud DataProc	EMR	HDInsight/ Azure Databrick	MapReduce Service / Data Lake Insight (DLI)	E-MapReduce	Big Data Service
	Streaming	Spark Streaming	Cloud Dataflow	Kinesis Analytics	Stream Analytics	Cloud Stream Service (CS)	Realtime Compute for Apache Flink	Data Flow
	Fast SQL	Hive	Google BigQuery	Athena	Data Lake Analytics	Data Lake Insight (DLI)	Data Lake Analytics	Autonomous Database
	Machine Learning	Spark MLLib	Cloud AI Platform	SageMaker	Azure ML	ModelArts	Machine Learning Platform For AI	Data Science
Visualisation	Dashboard	Tableau, Power BI etc.	Google Data Studio, Looker	QuickSight	Power BI	Data Lake Visualisation (DLV)	Quick BI, Data V	Oracle Analytic Platform

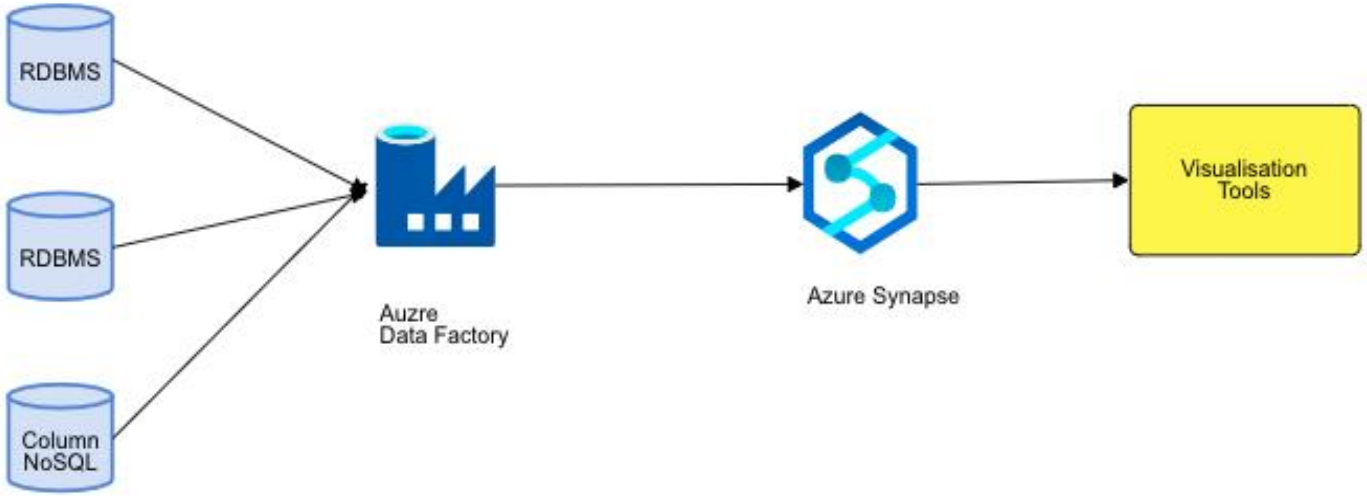
Data warehouse architecture on Azure



Data warehouse architecture on Azure



GROUP WORK 1

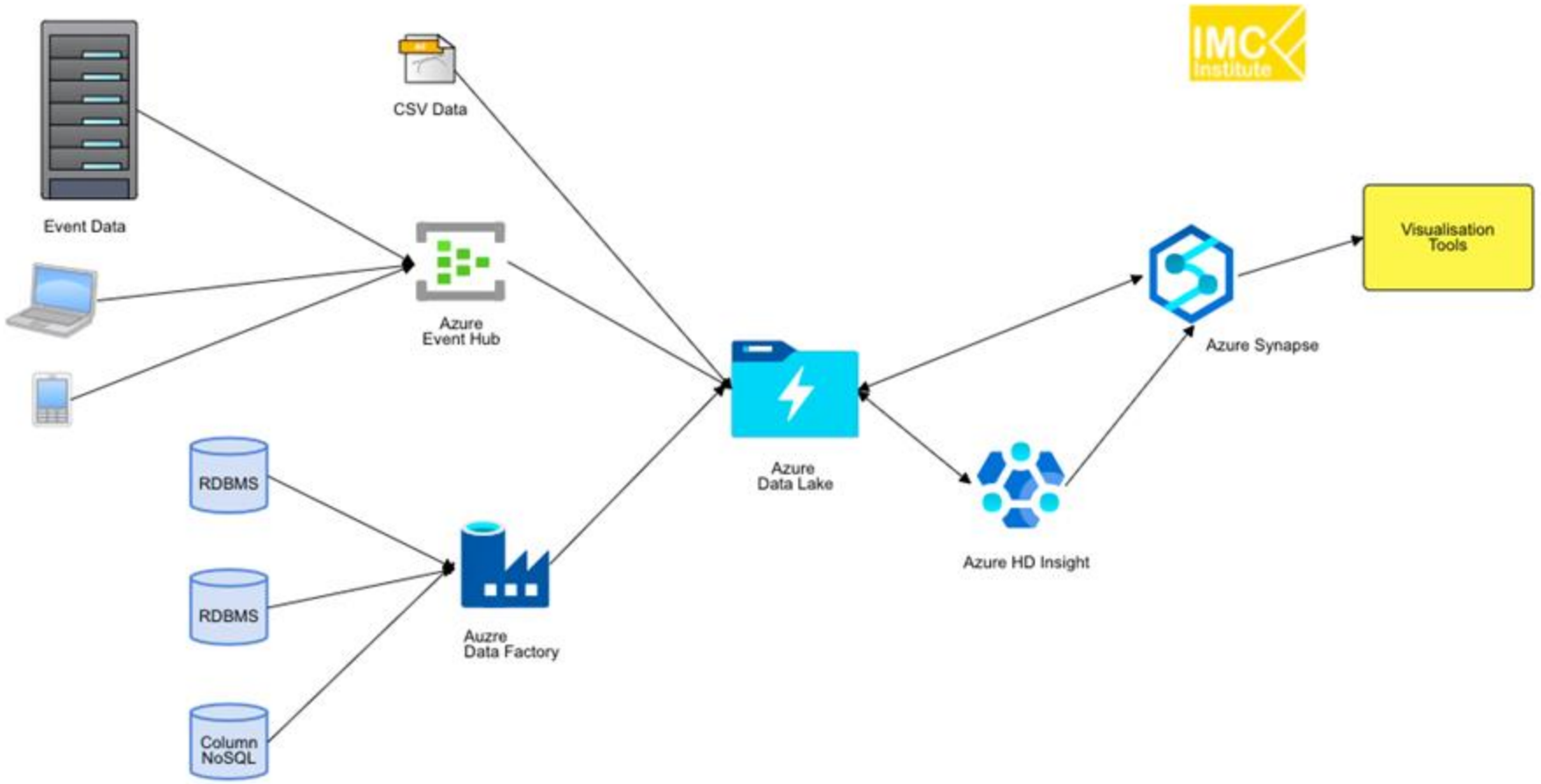


Compare it with

- GCP
- AWS

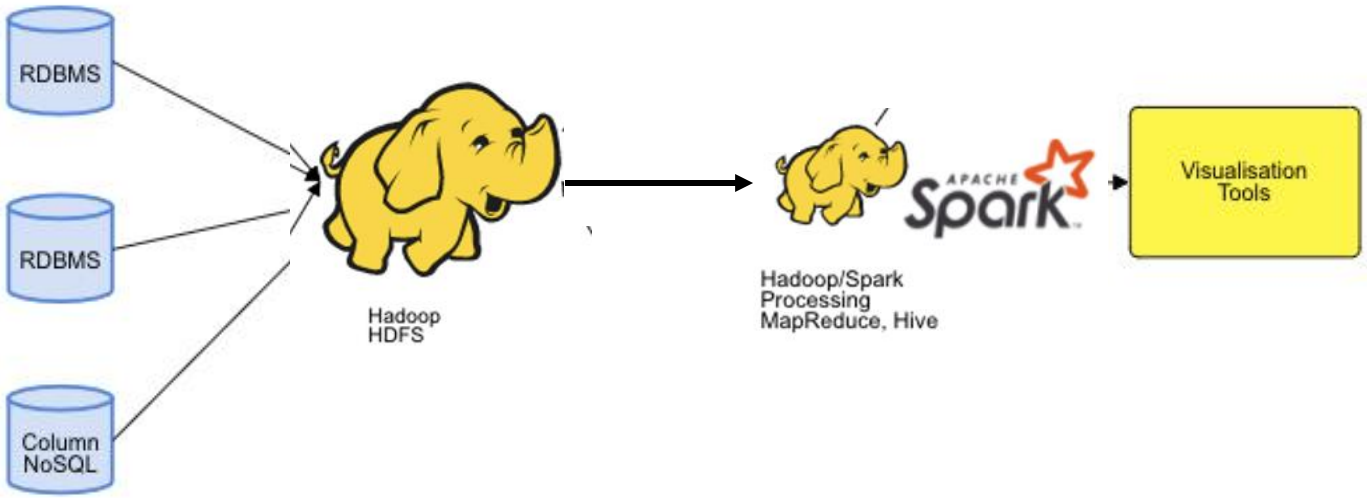
Mahidol University
Wisdom of the Land

GROUP WORK 2



Mahidol University
Wisdom of the Land

HDFS Architecture



GROUP WORK 1

1. Replace with other platforms

Compare it with

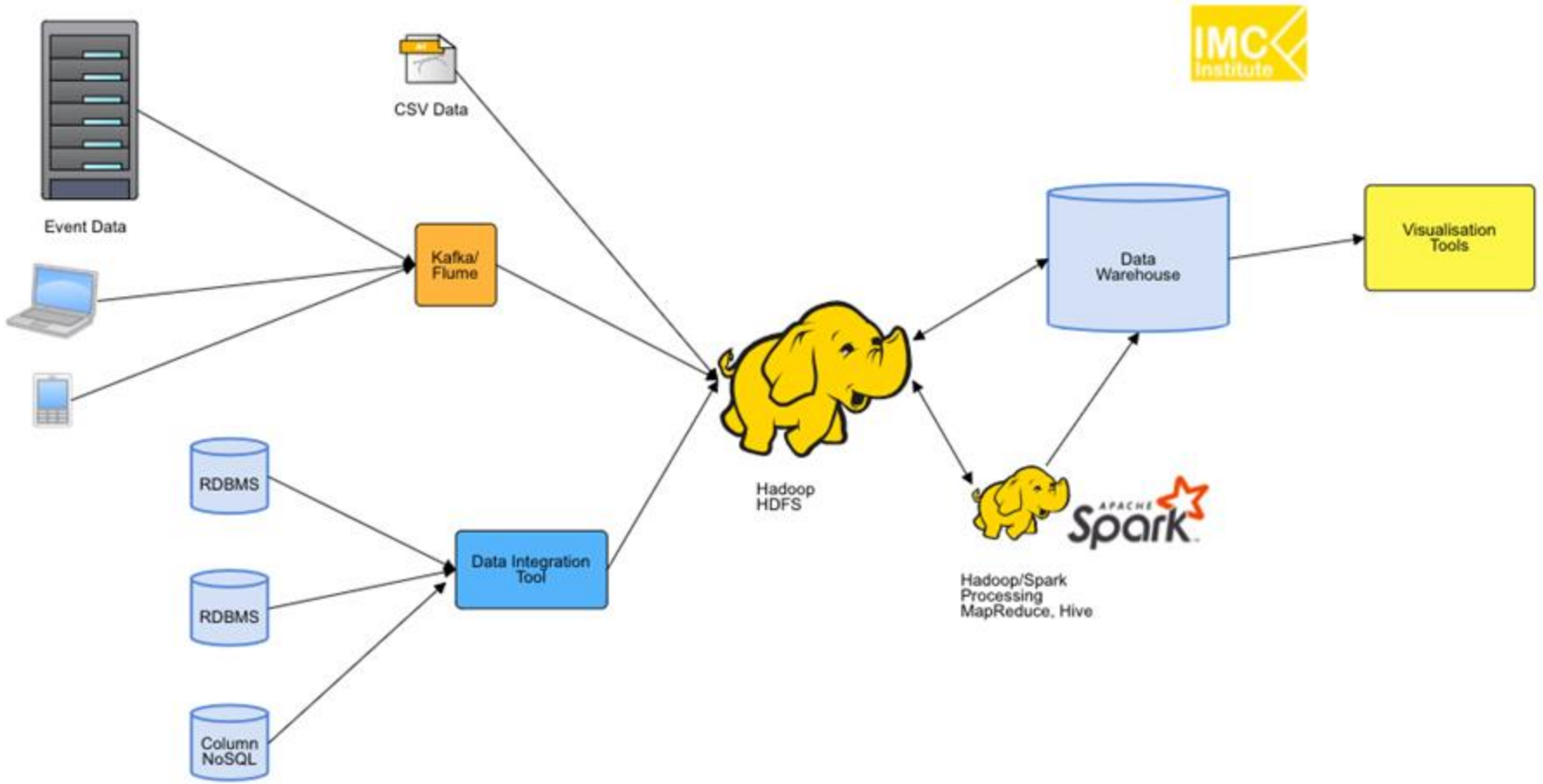
- GCP
- Azure
- AWS
- Huawei
- Alibaba
- Oracle

2. Check the price (assume usage by yourself)

- GCP https://cloud.google.com/products/calculator?hl=en
- Azure https://azure.microsoft.com/en-us/pricing/calculator/
- AWS https://calculator.aws/#/
- Huawei Price calculator https://www.huaweicloud.com/intl/en-us/pricing/calculator.html#/ecs
- Alibaba https://www.alibabacloud.com/en/pricing-calculator?_p_lc=1#/
- Oracle https://www.oracle.com/cloud/pricing/

Mahidol University
Wisdom of the Land

Group work2



Mahidol University
Wisdom of the Land

Price Check

- GCP https://cloud.google.com/products/calculator?hl=en
- Azure https://azure.microsoft.com/en-us/pricing/calculator/
- AWS https://calculator.aws/#/
- Huawei Price calculator https://www.huaweicloud.com/intl/en-us/pricing/calculator.html#/ecs
- Alibaba https://www.alibabacloud.com/en/pricing-calculator?_p_lc=1#/
- Oracle https://www.oracle.com/cloud/pricing/

Mahidol University
Wisdom of the Land

Issue with Big data on cloud

- Vendor lock-in
- Latency
- Data privacy on cloud
- Regulation/Compliance

Mahidol University
Wisdom of the Land

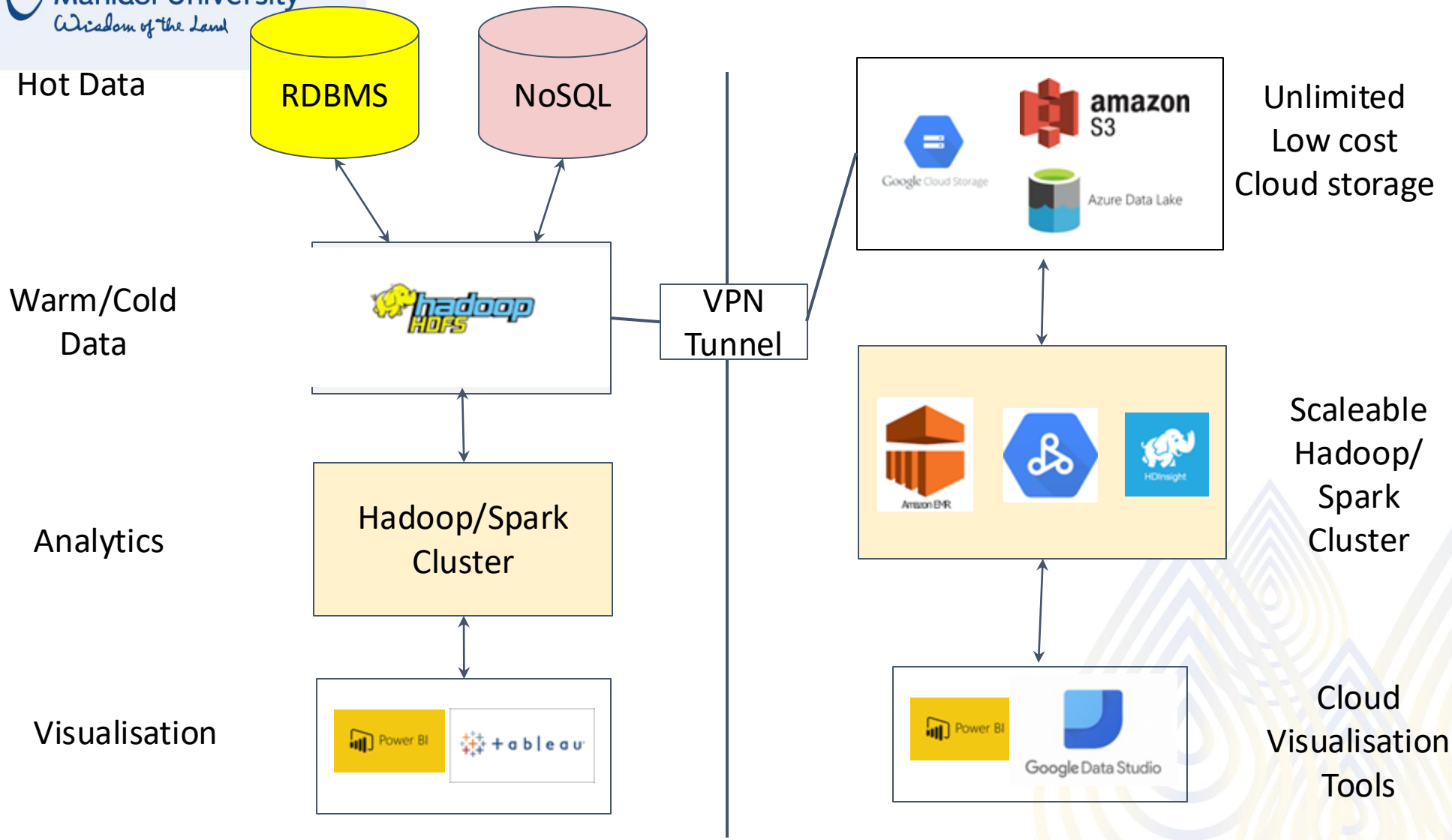
Hybrid & Multi Cloud Model

- Using both on-premise & public cloud services
- On-promise storage and processing for sensitive data
- Cloud storage for high scale data and also large archive data
- Scalable CPU for Hadoop as a services
- Hadoop on-premise distribution & Big data on cloud trend towards multi cloud model.

Mahidol University
Wisdom of the Land

Recommendation

- Always start with big data on public cloud
- Cloud storage as a main data lake
- Process large scale data using public cloud services such as Hadoop as a Service
- Small on-premise big data cluster for sensitive data and local processing.


Hot Data
RDBMS
NoSQL
Warm/Cold Data
hadoop HDFS
VPN Tunnel
Analytics
Hadoop/Spark Cluster
Visualisation
Power BI
+ableau
On-Premise
amazon S3
Google Cloud Storage
Azure Data Lake
Unlimited Low cost Cloud storage
Amazon EMR
HDInsight
Scaleable Hadoop/ Spark Cluster
Power BI
Google Data Studio
Cloud Visualisation Tools
Public Cloud

Mahidol University
Wisdom of the Land

HDFS vs. Cloud Storage: Pros, cons and migration tips

https://cloud.google.com/blog/products/storage-data-transfer/hdfs-vs-cloud-storage-pros-cons-and-migration-tips