Design Project

[ISSS612] Big Data: Tools and Techniques

DemeterAI

Team

Colin Jiang Kelin (colinjiang.2021@mitb.smu.edu.sg)
Khoo Kian Sim (kskhoo.2021@mitb.smu.edu.sg)
Lim Wei Jie (wjlim.2022@mitb.smu.edu.sg)
Perry Chia Dun Li (perry.chia.2021@mitb.smu.edu.sg)
Tong Zi Heng (ziheng.tong.2021@mitb.smu.edu.sg)
Yeo Yi Xuan (yixuan.yeo.2021@mitb.smu.edu.sg)

Design Principles

Cost

Keep cost low by re-using current capabilities or utilizing open source software, and limiting new technologies to as few as possible without affecting performance

Also avoid hardware which is too costly

Scalability

Able to scale both horizontally and vertically, as well as cater to new storage and processing types when needed

Agility

Bringing data to the right person and at the right timing

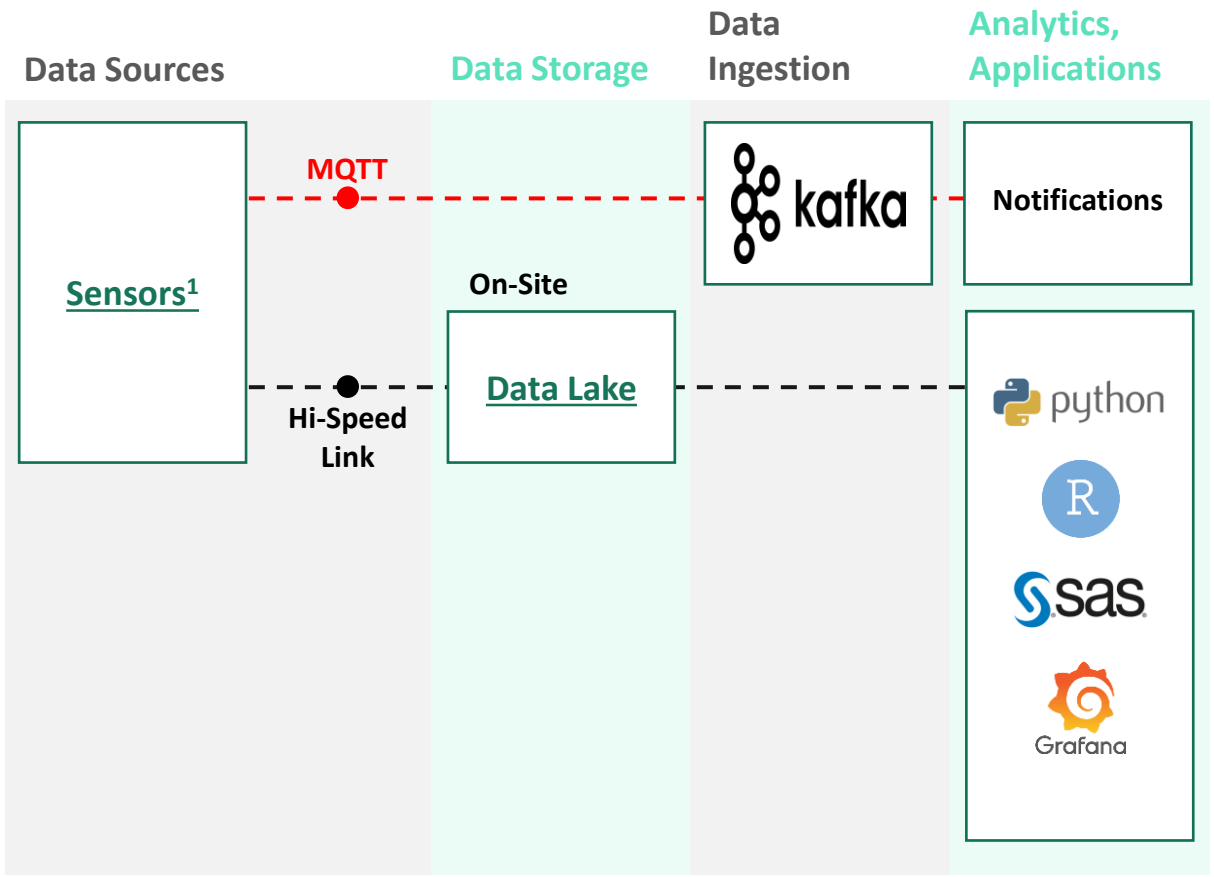
Integration

Prioritize high availability while maintaining eventual consistency to have single source of truth given inadequate broadband circumstances

Flexibility

Ease of integrating with existing technologies, as well as with new technologies when required in the future

Architecture View (As-is)



1. Sensors

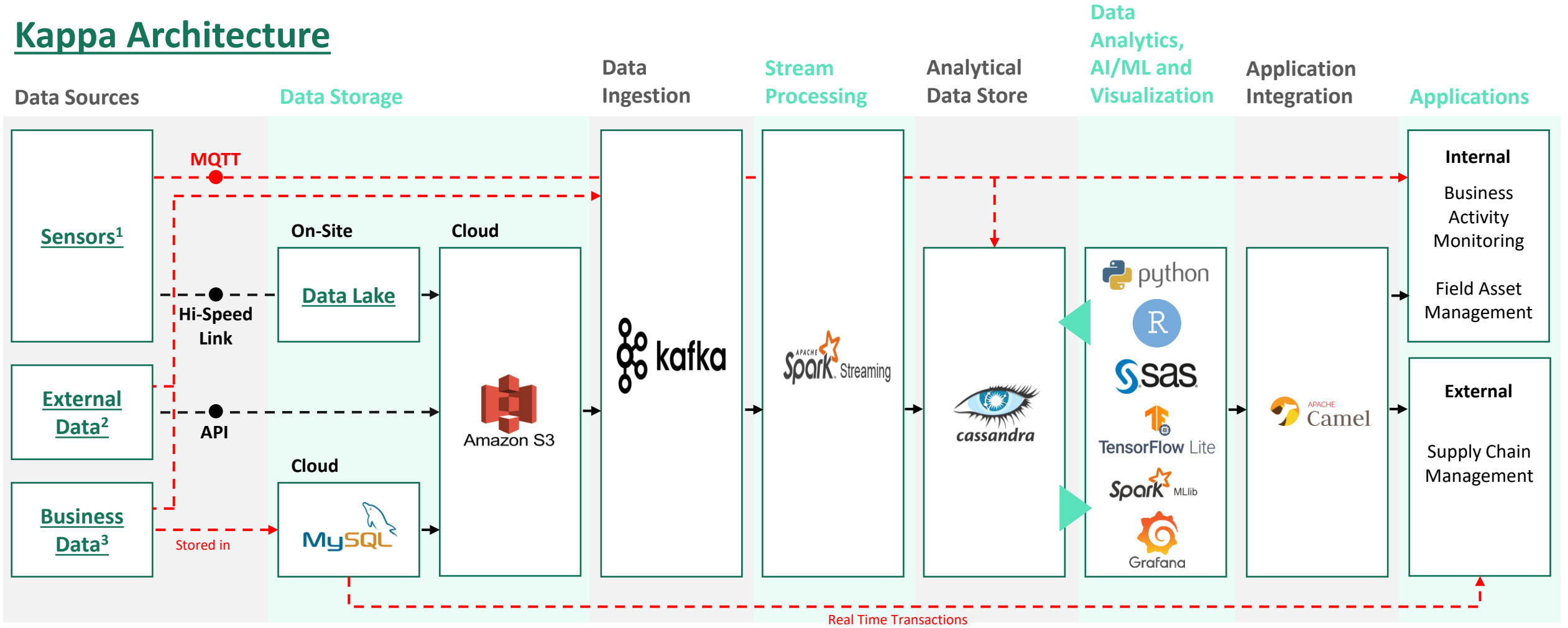
- IoT Sensor
- Weather Radar
- Camera

Real Time Data

Batch Data

Architecture View (To-Be)

Kappa Architecture



1. Sensors
- IoT Sensor
- Weather Radar
- Camera

2. External Data
- Futures Market
- Weather Forecast
- Satellite Imagery

3. Business Data
- Customer
- Supplier
- Inventory Level

Real Time Data

Batch Data

Technological Components

Tool	Why is it needed?	Existing	New
On-Site Data Lake	Data dump into on-site data lake before putting into cloud data lake (Amazon S3) Maintain data dumps for 5 years	✓	✓
MQTT	Standardized way to manage data flow for IoT system(s)	✓	✓
Kafka	Real-time stream-processing platform for various data sources including sensors, external data and business data	✓	✓
Python/R/SAS	Programming tool for data analytics	✓	✓
Grafana	Interactive visualization web application for analytics. Also send alerts based on pre-configured thresholds met for real-time data (business activity monitoring).	✓	✓
MySQL	Relational database management system for OLTP		✓
Amazon S3	Cloud storage service as data lake for easy scalability Maintain data dumps for 5 years		✓
SparkStreaming	Perform data transformations as needed		✓
Cassandra	NoSQL database management system which prioritises availability over consistency		✓
Tensorflow/TF lite	Deep learning application and deployment on edge devices		✓
Spark MLib	For machine learning algorithms such as classification, regression, decision trees and clustering		✓
Apache Camel	Integration framework to integrate data from multiple sources and exposing data as an API		✓

Assumptions

1. Restrictions in hardware upgrades, so only selected crucial events are streamed in real-time.

2. Employees are well versed in existing technology hence keeping most of the existing data analytics tools to minimize staff retraining.

3. Intention is to scale up significantly, so traffic will be high. Running on self managed open-source technology will aid in saving cost in the long run.