

ISSS612 Big Date: Tools and Techniques Design Project



Group 8

Vincent CHAN (vincentchan.2023@engd.smu.edu.sg)

Ilansurya ILANCHEZHIAN (ilansurya.i.2023@mitb.smu.edu.sg)

Neel Ketan MODHA (neelmodha.2022@mitb.smu.edu.sg)

MANSHARAMANI Harita Laveen (harital.m.2023@mitb.smu.edu.sg)

Kousika VELRAJ (kousika.v.2023@mitb.smu.edu.sg)

Naga Sivani BARLAPUDI (nagas.b.2023@mitb.smu.edu.sg)

Key Design Principles

Key Considerations:

- **Data as the Competitive Edge to Make Farm Cost Efficient**
 - Use data to improve crop yield. Keep data secured as a hack will impact yield/quality.
 - Use data to project yield and cash flow as they assure investors, govt (for grants) and banks (for lower interest loans¹).
- **Scalable and Agile System**
 - Architecture must be scalable and adaptive to variety of crop to keep up with changing market demands
- **Integrative Design with End-Users in Mind**
 - Maximise use of data integration and process automation to reduce repetitive work

Key Data Streams:

Environment Control

Pest Control

Pest Detection

Nutrients
(Potassium, ,
Sterilization (UV,
Ozone)

Detect nutrients
deficiency (Image
Sensors (20), TIFF
5MB, 1000/day

Acoustic Sensors
(2000), CSV,
1KB/sec

Light/Gas Sensors
(1000, 1000),
CSV, 1KB/30min

- Continuous adjustments with new farming methods and insights

Farm Automation

Sensor and
utilities data

Crop Monitoring

Notifications &
Up-to-Date
status

Telemetry (TIFF 5MB,
1000/day),
Actuators/Utility
(2MB/day/unit)

Video System (260),
1080p (80 GB/day
for video storage).

Crop Harvest
info(CSV, daily)

- Largely fixed due to building infrastructure design and installations

Business Operation

Inventory
(seeds,
fertilizers)

Demand
Forecast

Sales &
Channel
Performance

Warehouse
mgmt. sys
(CSV, daily)

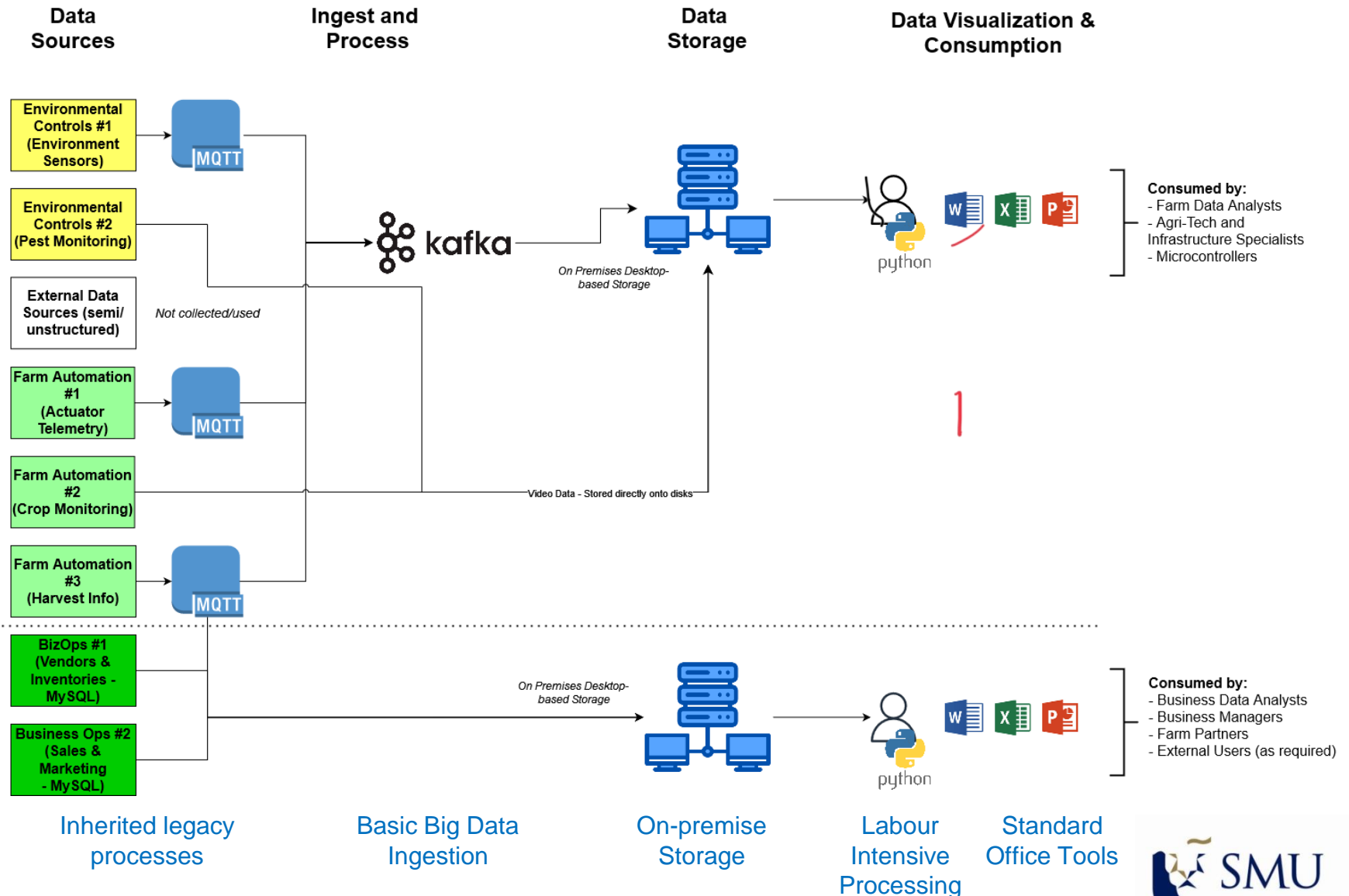
Sale Dept
(CSV, daily)

Sales Dept
(CSV, daily)

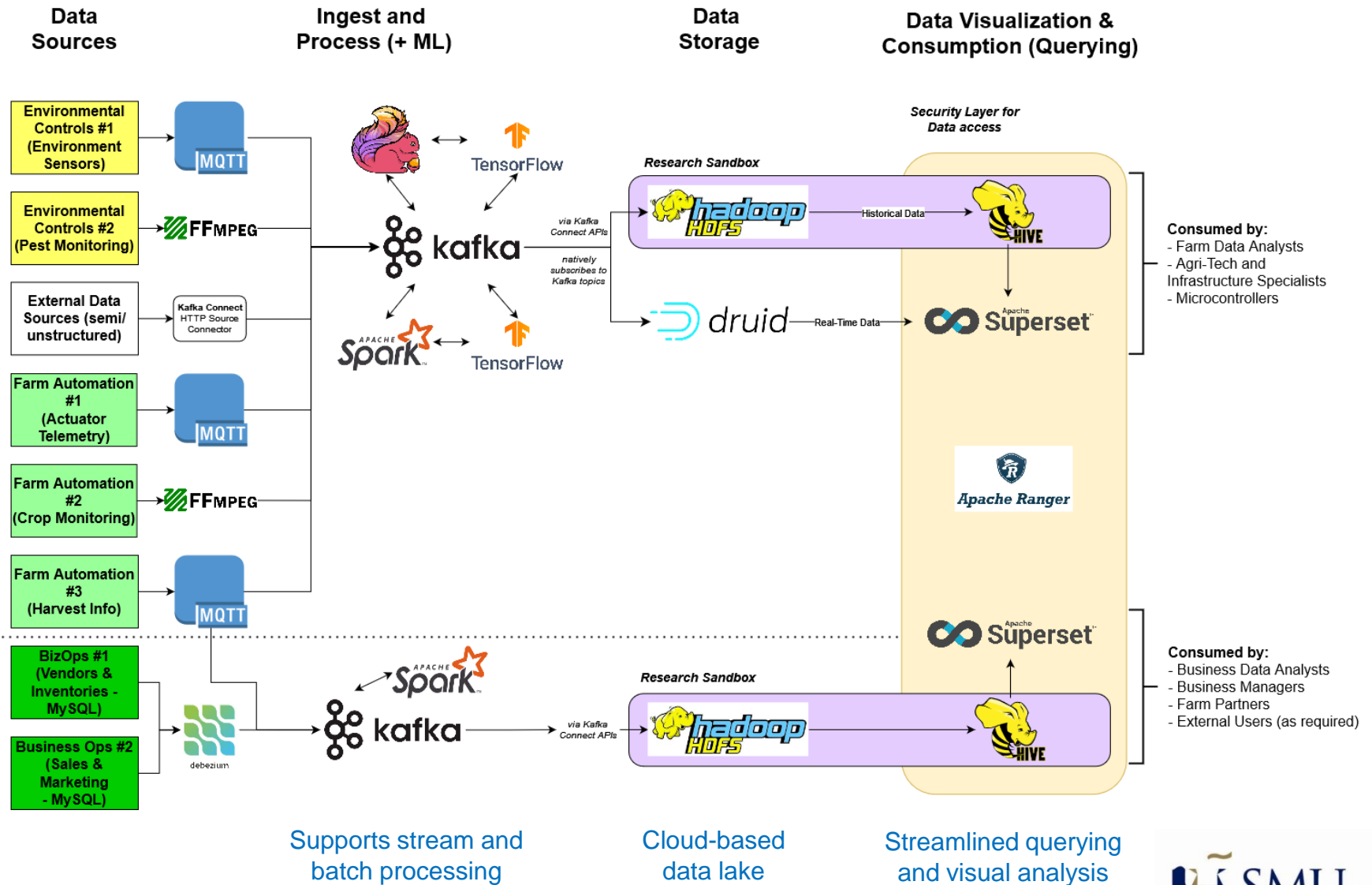
- Consider non-tech users and external stakeholder (banks, vendors, sale channels, direct marketing to customers)

[1] <https://www.forbes.com/sites/forbestechcouncil/2022/12/22/resolving-the-water-wars-the-4-key-drivers-of-indoor-vertical-farming-success/>

Architecture (Start State)










Architecture (Proposed)



Technology Used (Ingest & Process)

Present in

Component	Old	New	Purpose and Benefits
	✓	✓	<ul style="list-style-type: none"> Connector between Environmental Sensors/Farm Automation Telemetry data and Apache Kafka – Acts as a publisher of data to Kafka
	✗	✓	<ul style="list-style-type: none"> Connector between Farm Automation Video data and Apache Kafka – Acts as a publisher of data to Kafka
	✗	✓	<ul style="list-style-type: none"> A MySQL Connector that provides connectivity to the MySQL server for the client's business operations applications (e.g. SAP) – Acts as a publisher of data to Kafka
	✓	✓	<ul style="list-style-type: none"> Streams data from all sources – Sensors/Videos/Business Operations Apps to Spark/Flink for further processing, and then to dedicated data stores
	✗	✓	<ul style="list-style-type: none"> Processes environmental sensor, farm automation telemetry, and video data, in a real-time fashion to identify points where alerts may be required
	✗	✓	<ul style="list-style-type: none"> Processes ingested environmental sensor, farm automation telemetry, business operations and external data in a batch fashion, after which it publishes the processed data back to Kafka for storage subscribers
 TensorFlow	✗	✓	<ul style="list-style-type: none"> An ML library that integrates well with Apache Flink and Spark. Data being processed in Flink/Spark can be seamlessly run through ML models as part of the data pipeline (e.g., running video data through CV models to identify anomalies)

Technology Used (Store & Visualise)

Present in

Component

Old

New

Purpose and Benefits



- An analytical data store which can natively subscribe to and store real-time, time-series data from Kafka, used to power real-time dashboards on Apache Superset



- A general purpose data store which subscribes to data from Kafka (via Connect API) and used for longer term archival and batch storage of sensor, telemetry, video, business operations and external data.



- SQL-workbench for all data stores in HDFS, allows analysts to perform SQL-based analysis on longer-term historical data. Also connects to Apache Superset, allowing data to be visualized further



- Front-end data visualization tool, used by analysts to visualize data, partners to see real-time farm statistics/alerts, and business operations to manage and report on business metrics



- Enables monitoring and managing of data security and access across the Hadoop platform

Assumptions & Design Choices

- **Assumption 1:** External Data is assumed to be largely unstructured or semi-structured at best, due to the potential variations in schema (or lack thereof) from owners of these external data sources.
- **Assumption 2:** Business Operations Data is assumed to come from a MySQL database which is what the business application (e.g., SAP ERP system) uses as its primary database.
- **Assumption 3:** Storage for data is assumed to be residing on-premise, on disks in desktop-class computers, for the start state architecture
- **Design Choice 1:** The design is to leverage on cloud-based services as far as possible so to improve the ability to scale, enhance reliability and minimise upfront hardware investment.
- **Design Choice 2:** The technology stack for the architecture is carefully selected and analyzed to ensure fault tolerance and high availability.
- **Design Choice 3:** Farm data and Business Operations data is split into two data pipelines, to maintain separation between different functions of the business