Assignment 2: Project Status Report

Name	ID	Role	Background
Meheruna Alam	WQD190034	Solution Architect	Business Intelligence, Data Modelling & Data Architecture
Norzarifah Kamarauzaman	WQD190043	Solution Architect	Data Management (Geoscience) & Data Science
Owen Williams	WQD190037	Solution Architect	Software Development

1.0 PROJECT BACKGROUND

Recent advancement in technology has affecting energy industry in many ways [1], particularly resulting in a massive amount of data being acquired and processed across end-to-end value chains on a continuous basis through deployment of sensors, wireless transmission, network communication, and cloud computing technologies. Such a high volume, velocity, and variety of information assets termed as *Big Data* [2] must be managed efficiently to generate value from data. However, most organisations in the industry are facing challenges with *Dark Data* – the term coined by Gartner Inc. to describe information assets that organisations collect, process and store in regular business activities, yet are of little or no avail for other purposes such as business analytics and direct monetizing.

Clearly, due to disparate nature of data and technologies, business consumers often encounter difficulties when aggregating and coalescing data from heterogeneous database systems (including data-warehouses and data-lakes). Primary consequence of ineffective big data and knowledge management (i.e. when the organisations are unable to efficiently leverage its big data), is loss in competitive advantage which in turn comes at a cost. For example, at the upstream end of energy sector, delayed decision making in prospect evaluation during business acquisition may result to loss of investment, whereas at the downstream extremity, ineffective monitoring of energy production and utilisation will inevitably incur unnecessary cost to the consumers.

2.0 PROJECT PURPOSE

The main ***purpose*** of this project is to improve the processes of discovering, exploring and analyzing 'dark data' in an organisation, with the aim of maximizing derived value by facilitating best possible decision-making in timeliest manner. In identifying project scopes (hence solutions), we have adopted ***Design Thinking***; an iterative non-linear problem-solving model which seeks to understand users, challenge assumptions, redefine problems and create innovative solutions to prototype and test [3].

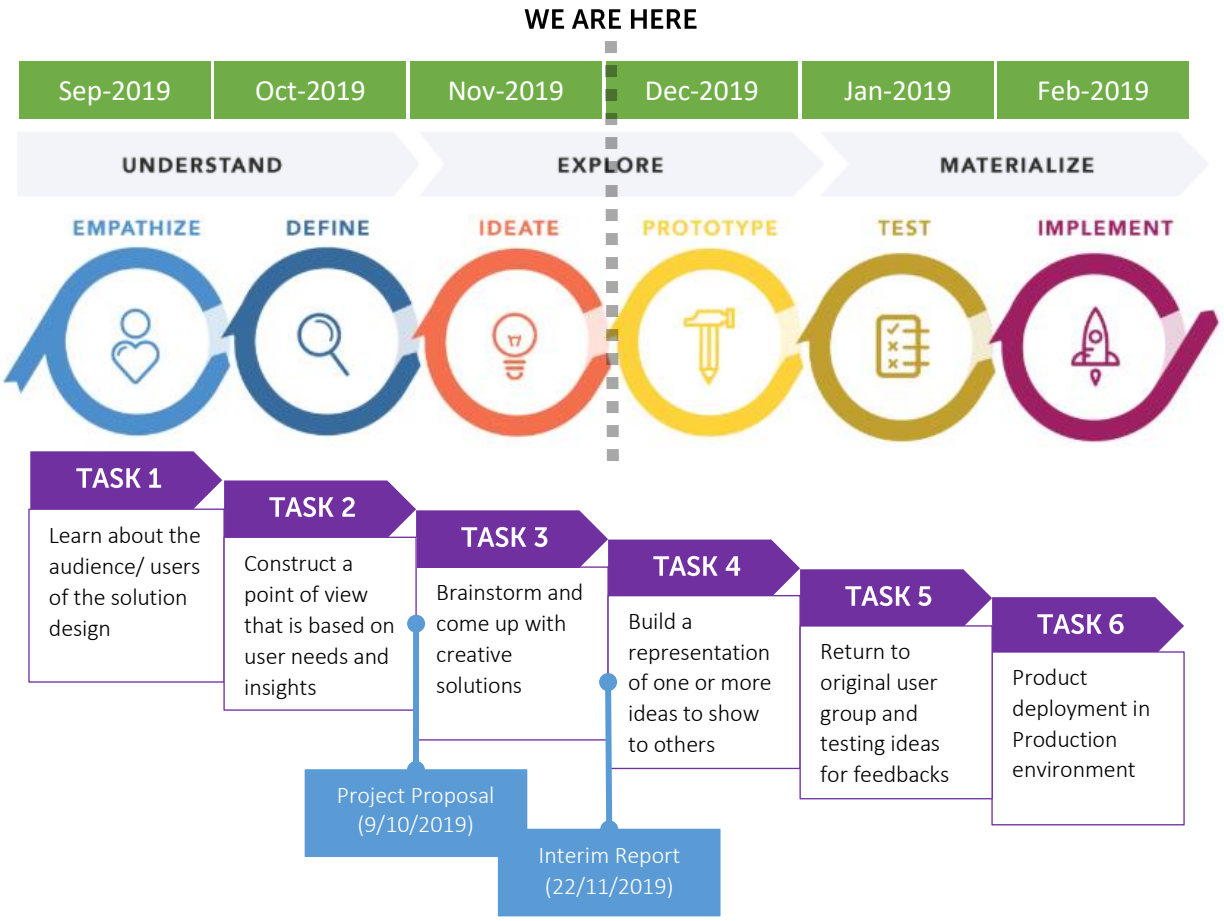
Upon sessions of brainstorming, we narrow down the **scope** of this project to designing a system architecture incorporating *Universal Query Language* and *Real-time Event Streaming System*, which aims to speed up Big Data management and data processing system to make an organization from reactive to proactive decision maker.

2.1 INTERIM REPORT PURPOSE

The purpose of this interim report is to update project progress to date, including all the efforts made to the present time in brainstorming the proposed solutions architecture that incorporate Universal Query Language and Real-time Event Streaming System.

3.0 PROJECT TIMELINE

As we are aligning our project milestones in accordance to the *Design Thinking* model, we are about halfway toward completing the project.



4.0 PROJECT PROGRESS

To date, the focus has been on carrying out literature review around relevant topics including: Case-studies of big data management in Energy Industry (production, distribution and monitoring) AND the two proposed solutions (universal query language and real-time event streaming).

4.1 LITERATURE REVIEW ON CASE STUDIES

- *Real-time Complex Event Processing and Analytics for Smart Grid:*

  Electric power networks are among the world's most complex human-made systems. The developing smart grid is an inherently complex system which is rapidly evolving in both definition and implementation. Deployment of advanced technologies within the electric utility sector and usage of state-of-the-art computing systems provides companies with innovative capabilities to forecast electricity demand, influence customer usage patterns, create demand response program, optimize unit commitment, and prevent power outages.

- *Make Your Oil and Gas Assets Smarter by Implementing Predictive Maintenance with Databricks:*

  Maintaining assets such as compressors is an extremely complex endeavor. Compressors are used extensively in offshore facilities, from small drilling rigs to deep-water platforms. These assets are located across the globe, and they generate terabytes of data daily. A failure for just one of these compressors costs millions of dollars of lost production per day. An important way to save time and money is to use machine learning to predict outages and issue maintenance work orders before the failure occurs.

- *Cloud-Based Software Platform for Big Data Analytics in Smart Grids:*

  There is a global effort to incorporate pervasive sensors, actuators and data networks into national power grids. This Smart Grid offers deep monitoring and controls, but needs advanced analytics over millions of data streams for efficient and reliable operational decisions.

4.2 LITERATURE REVIEW ON THE PROPOSED SOLUTIONS:

- *Universal Query Language*

  (To be provided)

- *Real-Time Event Streaming Architecture*

  Traditional data management model collects data from multiple sources and store it in RDBMS and Hadoop cluster. Subsequently, users need to process/query onto these storage systems to utilise the data for further action. However, such traditional method may lessen the value of the stored data as it could take huge processing time.

  Within the system itself, for all the streaming data captured from multiple sensors, or third party system are stored in messaging infrastructure – and users might just not knowing which part of data are useful within the storage hence potentially lead to dark data. Such slow data processing system affect proactive analysis for event streaming and in the case of Energy Industry, not having real-time intelligence can lead to safety issues, poor decisions, maintenance issues and can cost money.

  Therefore, to overcome this and to work with only with the potential data rather spending time and processes on unnecessary data we can introduce a system which will normalize, aggregate, cluster and filter to the necessary data and feed it to applications. Following figure is a draft proposal for such real-time event processing system.

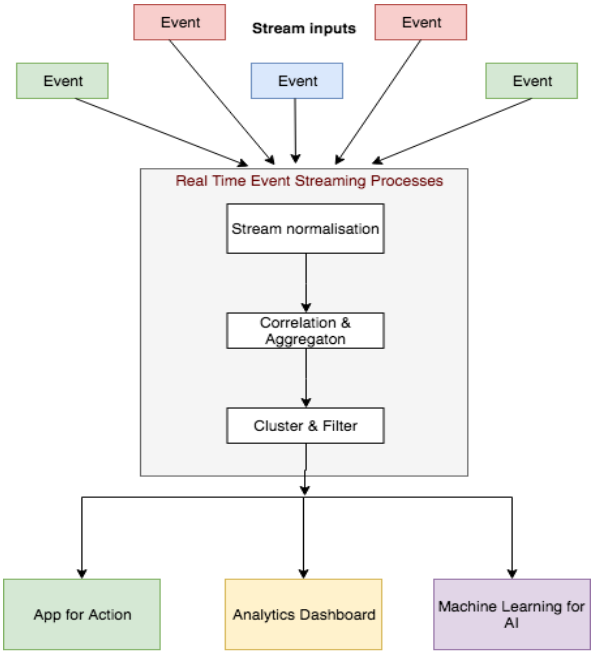


Figure: Real Time Event Streaming Process

In the first process, incoming stream input from multiple sources will be normalized to be feed to the next step. In next step, streaming data will be aggregated and correlated. In the last step, all the aggregated data will be clustered according to necessity of the data. In this step, all the dark data can be clustered or correlated in the same category. From this step users will get filtered data regardless unnecessary data to be fed into his system like Apps, Dashboard or ML system. Key feature of the architecture are fast access to live data, scalable, and fault tolerance.

5.0 REMAINING TASKS:

- The writers have a number of topic areas which still need to be explored or explored further.
- To further review the case-studies and provided more details analysis of their domain, encountered problems and proposed/chosen solutions.
- To do further literature review around universal query languages and real-time event systems i.e. available products and detailed comparison of their functionality

6.0 REFERENCES

[1] Madden, S. [2012]. From databases to Big Data. IEEE Internet Computing, vol. 16, pp. 4-6.

[2] Laney, D. [2001]. 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group: Stamford, CT, USA.

[3] https://www.interaction-design.org/literature/topics/design-thinking

[4] Big Data Management Canvas: A Reference Model for Value Creation from Data

[5] https://en.wikipedia.org/wiki/Event_stream_processing

[6] https://medium.com/stream-processing/stream-processing-101-from-sql-to-streaming-sql-in-10-minutes-5edcb10e56e9

[7] https://www.sciencedirect.com/science/article/pii/S1877050915029993

[8] https://databricks.com/blog/2018/07/19/make-your-oil-and-gas-assets-smarter-by-implementing-predictive-maintenance-with-databricks.html

[9] https://www.researchgate.net/profile/Yogesh_Simmhan/publication/260585847_Cloud-Based_Software_Platform_for_Big_Data_Analytics_in_Smart_Grids/links/599b9f04a6fdcc500349c9fb/Cloud-Based-Software-Platform-for-Big-Data-Analytics-in-Smart-Grids.pdf

[10] https://www.tibco.com/blog/2019/10/16/to-become-a-real-time-business-you-need-event-streaming-architecture/

[11] https://en.wikipedia.org/wiki/Apache_Kafka

[12] https://databricks.com/glossary/what-is-spark-streaming

[13] https://aws.amazon.com/marketplace/pp/Databricks-Inc-Databricks-Unified-Analytics-Platfo/B07K2NJKRW

[14] https://www.researchgate.net/profile/Yogesh_Simmhan/publication/260585847_Cloud-Based_Software_Platform_for_Big_Data_Analytics_in_Smart_Grids/links/599b9f04a6fdcc500349c9fb/Cloud-Based-Software-Platform-for-Big-Data-Analytics-in-Smart-Grids.pdf