Real Time Data Ingestion Platform

LF Energy TAC Project Proposal

General information

- **Name**
  Real Time Data Ingestion Platform
- **Mission statement**
  Easy access to high volume, historical and real time process data for analytics applications, engineers, and data scientists wherever they are.
- **Description (what it does, why it is valuable, origin and history)**
  RTDIP was built internally at Shell to process time series data from Assets in the Cloud into a common data lake so that it could be used for analytics, ML and AI use cases.

  After scanning the market for suitable products as well as working with consultancies in 2017, Shell deemed there to be nothing on the market that was able to ingest and process data at the scale required for their Predictive Maintenance Program. It was decided in 2018 that Shell would build a product that was able to stream their Asset data to the cloud and store it in a Data Lake for consumption by products and users.

  The product was built be Shell's data engineering group and extensively uses Apache Spark and Databricks to process streaming, real time data into the Delta format so that it can be easily consumed by data engineers, data scientists and business users.

  The project being open sourced provides:
  1. The Delta Ingestion engine used to process streaming data from streaming sources and files stored in cloud storage into Delta format. The data ingested is typically sourced from Pi Historians, OPC UA Servers, IoT Devices
  2. A Python SDK that enables data consumers to read and query raw, sampled, interpolated or time weighted averages of the data stored in Delta
  3. REST APIs that are wrappers for the Python SDK that enable developers in non-python languages to consume the data
  4. End User Documentation for using the SDK and Rest API

  SSIP(The Internal Shell name of RTDIP) provides data to both the C3.ai Platform where Machine Learning and AI production grade applications are built and operated, as well as to Digital Twins. Additionally, there are over 50 projects internally at Shell that consume SSIP Data for generating their insights. SSIP is deployed at every operated renewable Asset at Shell.

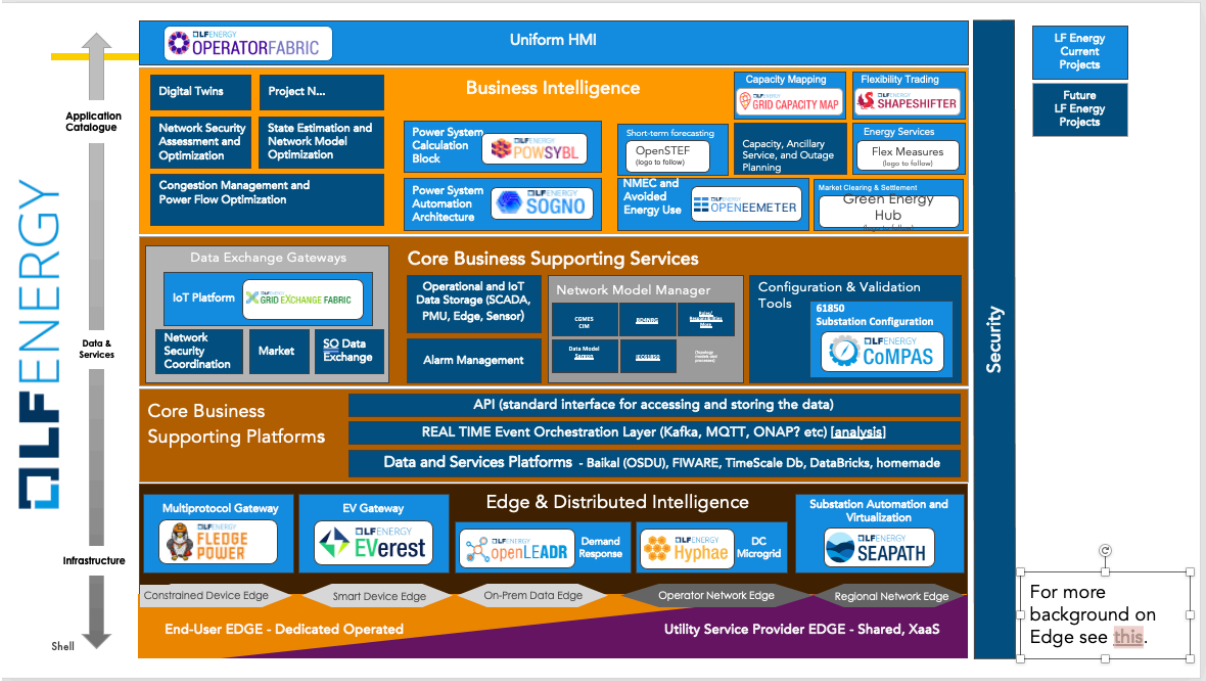
- **Is this a new project/working group/special interest group or an existing one?**
  This is an existing project within Shell. Development started in 2018 and continues to be funded internally by Shell as it is a fundamental service to the Enterprise.

The project is governed by a group by 8+ people internally at Shell, with a development and operations team of 14 people.

- **Current lead(s)**
  Bryce Bartmann
- **Sponsoring organization(s), along with any other key contributing individuals and/or organizations**
  Shell PLC
- **Detail any existing community infrastructure, including:**
  - **Github/GitLab, or other location where the code is hosted**
    https://github.com/sede-open/real-time-data-ingestion-platform
  - **Website and/or docs**
    https://sede-open.github.io/real-time-data-ingestion-platform/
  - **Communication channels (such as Mailing lists, Slack, IRC)**
    None exist yet publicly(only private)– will be created once Open Sourced
  - **Social Media Accounts**
    None exist yet – will be created once Open Sourced
- **Are there any specific infrastructure needs or requests outside of what is provided normally by LF Energy (please refer to the lifecycle for project benefits)? If so please detail them.**
  Possibly github, but Shell can host the code in their sede-open repository
  Additionally, if LF Energy has access to code quality scanning tools like Sonarcloud, we would want to take advantage of that.
- **Why would this be a good candidate for inclusion in LF Energy?**
  After discussion with LF Energy, we can see an opportunity to provide an open source, multi-cloud, time series data platform for the services that operate within LF Energy. This can support and incubate ML and AI type use cases for the Energy Industry
- **How would this benefit from inclusion in LF Energy?**
  Real Time Data is required in almost all industries, and extensively in the Energy Industry. The RTDIP can assist to standardise on how time series data is ingested and queried across the industry. Additionally, contributors from various sectors of the energy industry can help to build ingestion pipelines and time series queries for the various types of time series data that exists.
- **Provide a statement on alignment with the mission in the LF Energy charter.**
  SSIP has enabled Shell to bring digitalization to its Assets, whether mature or new. Its enabled Shell to embark on modernizing and digitally transforming how Shell operates assets today. By standardising on a common real time data platform, it will facilitate interoperability, automation, virtualization, flexibility and digital orchestration. SSIP in its private version has achieved this, and this can be used to assist with the same statement found in the LF Energy charter.
- **What specific need does this project/working group/special interest group address?**
- It addresses the requirement to ingest and query time series data in a modern tech stack
  - Apache Spark is used to ingest data into a Delta Lake that can integrate with Unity Catalog for Security and searching of data
  - Queries that can be executed via ODBC, Rest API or an SDK
  - Enables ML, Optimisation, Analytics use cases for LF Energy that can be done using open source products like scikit learn, MLFlow etc

- **Describe how this project/working group/special interest group impacts the energy industry.**
  The internal version of this product has enabled over 50+ projects at Shell to consume and use this data. It enables use cases such as:
    - Predictive Maintenance
    - Optimization of processes
    - Forecasting
    - Digital Twins

  These are all applicable to the energy industry
- **Describe how this project/working group/special interest group intersects with other LF Energy projects/working groups/special interest groups.**
  As per the slide below on LF Energy, there is a “future” plan to provide Operations and IoT Data Storage. RTDIP would fit this bill perfectly and would be best suited to work with any of LF Energy’s edge or IoT Products(Grid Exchange Fabric). OPC-AE(Alarms and Events) data is also possible to store in RTDIP. Additionally, given that RTDIP is based on a Delta Lakehouse architecture, it is possible to extend the data store to more than just time series data.



- **Who are the potential benefactors of this project/working group/special interest group?**
  Process Engineers are typically significant benefactors of access to process time series data.
  Additionally, this can be used by operators of process control domains or those responsible for monitoring equipment.
- **What other organizations in the world should be interested in this project/working group/special interest group?**
  Any organisation that currently operates assets or equipment where sensors and measurements generate time series data could be interested in using RTDIP. It is built using open source software with a view to standardise on how to ingest time series

data, how to model the data and common query patterns to be used on it. Given its mostly cloud agnostic, it would appeal to organisations on any of the major clouds.

- **Plan for growing in maturity if accepted within LF Energy**
  Shell is committed to the future of SSIP with a long term plan to continue to invest and mature the product internally. This commitment would move to focussing on the open source version

Technical Projects Questions

- **Project license**
  Apache 2.0
- **Is the project's code available now? If so provide a link to the code location.**
  Still being prepared for imminent release, but can be found at https://github.com/sede-open/real-time-data-ingestion-platform
- **Does this project have ongoing public (or private) technical meetings?**
  SSIP is managed by a Natural Delivery Team internally at Shell.
- **Do this project's community venues have a code of conduct? If so, what is it?**
  There is an internal code of conduct, but we are updating this to be the LF Energy Code of Conduct
- **Describe the project's leadership team and decision-making process.**
  The Natural Delivery Team(NDT) that leads the project comprises of:
  - Product Owner
  - Product Manager
  - Operations Manager
  - Security Architect
  - Solution Architect
  - Resource Manager

  Decision making processes are currently managed internally, where the NDT meet every 3 weeks to discuss any items that require decisions our outcomes.
- **Does this project have public governance (more than just one organization)?**
  No
- **Does this project have a development schedule and/or release schedule?**
  SSIP has been built internally since 2018 following scrum methodology operating on 2 week iteration cycles. Releases follow scrum cycles.
- **Does this project have dependencies on other open source projects? Which ones?**
  - Apache Spark
  - Delta.io
  - Databricks-sql-connector
  - FastAPI
  - Pyodbc
  - Turbodbc
  - Mkdocs
- **Describe the project's documentation.**
  Documentation has been built and deployed using Material for Mkdocs. This can be found at this link https://sede-open.github.io/real-time-data-ingestion-platform/ This documentation is targeted at consumers/users of the RTDIP SDK and Rest API.

Additionally, the repo is maintained to contain the required files(LICENSE, CONTRIBUTING, CODE_OF_CONDUCT etc) with developer guidelines on how to get started with the project

- **Describe any trademarks associated with the project.**
  There are no trademarks related to this
- **Do you have a project roadmap? please attach [Are this project's roadmap and meeting minutes public posted?]**
  An example of our 2022 roadmap can be seen here https://sede-open.github.io/real-time-data-ingestion-platform/roadmap/yearly-roadmaps/2022-development-roadmap/

  We will publish the 2023 in the coming weeks when our internal funding has been confirmed.
- **Does this project have a legal entity and/or registered trademarks?**
  Not that I am aware
- **Has this project been announced or promoted in any press?**
  Shell has regularly talked about SSIP and its successes at conferences(C3 Transform, Data and AI Summit. It has never been discussed or promoted from an open source perspective.
- **Does this project compete with other open source projects or commercial products?**
  Time Series storage is available in other products.
  Open Source examples are:
    - MongoDB
    - Postgres
    - Timescale
    - InfluxDB

  Commercial Options:
    - Azure Data Explorer
    - AWS Timestream