Data Technology Landscape

|  |  |
| --- | --- |
| Apache Spark | **unified analytics engine** -> query from source, perform data transformation -> put into another source\* Core:**\* SQL + DF****\* MLib:****\* Streaming:****\* GraphX:** |
| Apache Flink |  |
| Delta Lake |  |
| Apache Hive |  |
| Apache Hadoop |  |
| Apache Samza |  |
| Hazelcast JET |  |
| Apache Airflow |  |
| Apache Beam |  |
| Apache NIFI |  |
| Apache Kafka |  |
| Databrick MLFlow |  |
| DatabrickKoalas |  |
| Dask |  |
| Ray |  |
| Modin |  |
| XArray |  |
| Vaex |  |
|  |  |
|  |  |

Decision Point:  
\* Maintainable  
\* Not overkill  
\* Efficiency Enough -> definition of enough  
 \* Latency  
 \* Memory usage  
 \* Cost of usage  
 \* Setup efforts  
 \* Support at least Python/Java/Scala

\* Databrick is the direct competitor to our Project -> How ever what if we only specific into FX Market.

\* Basic Structure:

\* Legacy: Datasource ---> Crawler ---dump---> Datalake ---ETL---> Data Warehouse

\* Streaming: Datastream ---> Receiver ----> DStream (bunch of RDD) --->