

* Above Architecture depicts how we capture data from multiple source as event streams and write data into a Datalake and ingest certain data into a datawarehouse for data modeling to be used for BI/Reporting tools
* We are getting weather events stream and reading that data using Kafka producer.  
  Data will be pushed and created as Topics and will be fetched by consumer.
* Spark uses the data from kafka and creates Data frame to run the data frame we can use spark engine!  
  Once we get the output we can push the same to S3 (either in csv or Parquet )
* Table will be created , then push it to Data Warehouse (eg: snowflake ,redshift) for Data modeling Erwin is preferred.

Benefits and Drawbacks of Each tools:

Related to project:

**Kafka** –

THREE MAIN FEATURES OF APACHE KAFKA  
  
1. MESSAGING SYSTEM  
  
Highly Scalable, Fault Tolerant, Resilient, Distributed, Enterprise grade Publishing/Subscribing messaging system. It communicates through asynchronous calls resulting in an efficient system  
  
2. STORAGE SYSTEM  
  
A fault tolerant, distributed, highly efficient, scalable, durable and efficient storage system that is highly suitable for larger distributed enterprise level softwares.  
  
3. STREAMING PLATFORM  
  
On the fly and real time streaming and analysis of data. Can deal with 100s of different data sources such as Logs, Tables, Files, Messaging, Queues, Result Sets and others.  
  
**APACHE KAFKA CORE COMPONENTS**  
  
1. PRODUCER & CONSUMER API  
  
Producers and Consumers deal with creating data and reading that data. Both APIs are core components of a Kafka System. Producers create TOPICS that are read by other Consumers. TOPICS can be in different formats such as JSON.  
  
2. KAFKA CONNECTORS  
  
Apache Kafka integrates with more than 200 different technologies and platforms. Those platforms can read data from Kafka queues and store it in their DBMS or they can create data and place it in Kafka queues. These connectors can be producers or consumers. Organizations can create their own connectors as per their business need.  
  
Connectors are either SOURCE (Cloud DBMS, RDBMS, NoSQL DBMS) or SINK (Monitoring, Logging Tools, Elasticsearch, or Batch Systems - Apache Hadoop)  
  
3. STREAM API  
  
In BigData and Enterprise Systems data is often in huge volume and Kafka helps those systems achieve high performance with large amounts of data by providing transforming streams such as data parallelism, distributed coordination and concurrent data ingestion.

KAFKA ADVANTAGES  
  
1. Low Latency  
2. High Output  
3. Fault Tolerant System  
4. Distributed System  
5. Real Time Handling  
6. Excellence for Automation  
7. Works well with Batch Systems  
8. Scalable System  
9. Useful in Data Pipelines & DataOps  
10. Can integrate with 200+ Data tools  
11. Useful for many types of data work (Streaming, Parsing, Storing, Communicating, Analyzing.....)  
  
KAFKA DISADVANTAGES  
  
1. Difficult Monitoring and Tracing  
2. Reduced Performance  
4. Sometimes clumsy  
5. Request/Reply communication absent  
6. Point to Point communication absent  
7. Not easy to debug and trace

**Spark**-

**Benefits:**

* Parallel Processing
* Easy to work with Dataframes using python and SQL
* In memory processing
* Spark cluster is reliable
* Connection to AWS S3 is very easy

**DrawBacks:**

* No real time processing
* File Management System – only if you are really dependent on spark storage , if not you can manage it! (we are using s3)

**Airflow**

**Benefits**

* Jinja Templating
* Reusable Parameterized config files
* Airflow is open source community support is awesome
* setup is easy

**Drawbacks**:

* Configuration overload right from the start + hard to use locally.
* Setting up Airflow architecture for production is NOT easy.
* Lack of data sharing between tasks encourages not-atomic tasks

Moving past Airflow: Why Dagster is the next-generation data orchestrator.  
Below reference link as deep dive comparison of upcoming Dagster with Airflow:  
  
<https://dagster.io/blog/dagster-airflow>

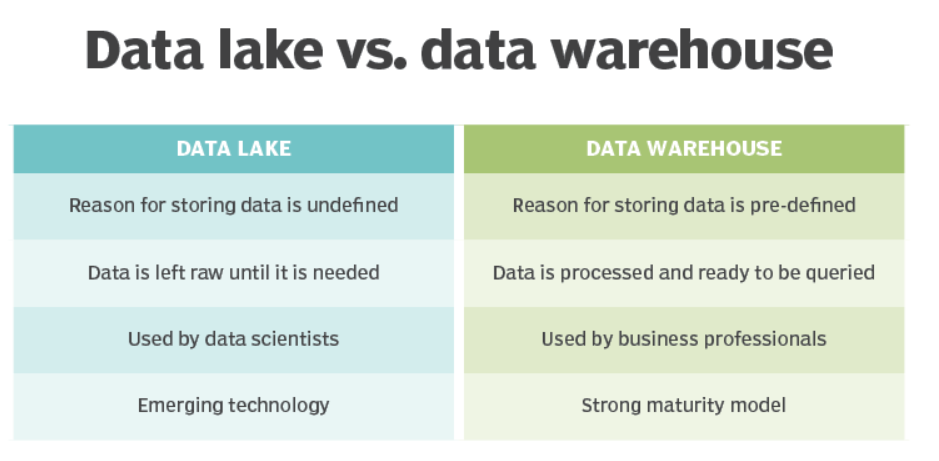
**Data Lake and Data Warehouse:**

A Data lake is a strategy that can be applied to multiple technologies. For example, the physical architecture of a data lake using Hadoop might differ from that of data lake using Amazon Simple Storage Service.  
  
**Benefits**:

* The ability of developers and [data scientists](https://searchenterpriseai.techtarget.com/definition/data-scientist) to easily configure a given data model, application, or query on the fly.
* Data lakes are cheap to implement because most technologies used to manage them are open source (i.e., Hadoop) and can be installed on low-cost hardware
* Agility allows for a variety of different analytics methods to interpret data, including big data analytics, real-time analytics, [machine learning](https://searchenterpriseai.techtarget.com/definition/machine-learning-ML) and SQL queries.

**Drawbacks**:

* Data lake is that it might turn into a data swamp, or data graveyard. If an organization practices poor data governance and management, it may lose track of the data that exists in the lake, even as more pours in.



**Monitoring and Reporting:**  
  
**Tableau** is a very effective tool to create interactive data visualizations

|  |  |
| --- | --- |
| **Tableau** | **Power BI** |
| Tableau BI can handle a huge volume of data with better performance. | Power BI can handle a limited volume of data. |
| Tableau platform is known for its data visualization functionality. The users can use 24 different types of visualizations in Tableau. | PowerBi offers numerous data points to offer data visualization. lt is offering more than 3500 data points for drilling down the dataset. |
| Tableau has excellent customer support. It has a large community forum for discussions. | Power Bl provides limited customer support to its users with a free Power Bl account. However, the paid version users will get faster support compared with the free version. |
| Tableau works best when there is a vast data in the cloud. | Power Bl doesn't work better with a massive amount of data |
| Analysts and experienced- users mostly use for their analytics purposes. | It is used by both naive and experienced users. |
| Tableau is a little difficult. | Power BI Interface is very easy to learn. |
| The information can be stored by using the Tableau server. | Power BI concentrates more on reporting and analytical modeling but not for storing the data. |
| Suitable for medium & Large type of Organization. | Suitable for Small, medium & large type of organization. |
| Tableau can connect to numerous data sources. | Power BI connects limited data sources while increasing its data source connectors in monthly updates. |
| Embedding report is a real-time challenge in Tableau | Embedding report is easy with Power BI. |