Apache Pyspark:

Big data processing is facilitated by Apache Pyspark, a potent open-source distributed computing platform. It provides a high-level Python development API on top of the Apache Spark platform. Large dataset processing, intricate data changes, machine learning, and graph processing are all excellent. applications for Pyspark. Scalability, quickness. That's why I choose it.

Apache Airflow:

Apache Airflow is an open-source platform that is used to programmatic ally author, schedule, and monitor workflows.It enables users to design intricate workflows that can be directly or routinely triggered. For orchestrating data pipelines and managing intricate ETL workflows, Airflow is perfect. Its adaptability, simplicity of use, and expansion are advantages. Therefore I choose this.

Apache Kafka:

Apache Kafka is an open-source distributed streaming platform that is used to build real-time data pipelines and streaming applications and It permits the development of real-time data sources that are capable of real-time processing and analysis. That is the reason I choose this.