**Introduction** At my current job, I am working on sophisticated data engineering projects aimed at analyzing customer behavior and enhancing targeted marketing strategies. These projects leverage a range of big data tools and technologies to manage, process, and analyze large volumes of data collected from various sources. Our architecture is designed to handle both batch and real-time data, ensuring timely and actionable insights for the business.

**High-Level System Architecture** Our system architecture includes:

* **Apache NiFi:** Data ingestion and initial processing.
* **Apache Kafka:** Real-time data streaming.
* **Apache Spark:** ETL (Extract, Transform, Load) processes.
* **HDFS (Hadoop Distributed File System):** Data storage.
* **Apache Ozone:** Additional storage.
* **Orchestration and Monitoring Tools:** Apache Airflow, Apache Ambari, Oozie, the ELK Stack, Apache Zookeeper, and Apache Falcon.

**Data Ingestion**

* **Sources:** Web logs, mobile app interactions, IoT sensors, and external FTP clients.
* **NiFi:** Configured to collect raw data, perform initial processing tasks (filtering, routing, format conversion), and publish real-time data to Kafka.
* **Kafka:** Streams real-time data collected by NiFi, providing a reliable and scalable way to handle real-time data streams.

**ETL Process with Apache Spark**

* **Raw Data Storage:** Data ingested by NiFi is stored in HDFS, organized by date and source. Real-time data from Kafka is also stored in HDFS for historical analysis.
* **Transformations:** Apache Spark performs ETL processes at regular intervals, reading raw data from HDFS, performing necessary transformations, and writing processed data back to HDFS. For real-time processing, Spark Streaming reads data from Kafka, processes it in real-time, and writes results to HDFS or other storage solutions.

**Data Pipeline Management**

* **NiFi and Kafka:** Manage the flow of transformed data, including additional processing steps such as data enrichment, integration with external systems, and routing to different storage solutions based on business requirements.

**Scheduling and Orchestration**

* **Apache Airflow:** Schedules and orchestrates ETL jobs and other data processing workflows. Airflow DAGs (Directed Acyclic Graphs) define the sequence of tasks and manage dependencies between them, ensuring reliable and timely processing.

**CI/CD Pipeline**

* **Version Control with Git:** Manages all code, including NiFi configurations, Spark jobs, and Airflow DAGs, allowing for tracking changes, effective collaboration, and ensuring code quality.
* **CI/CD with GitHub Actions:** Automates the build, test, and deployment processes, ensuring new code changes are thoroughly tested and deployed to the production environment without manual intervention.

**Monitoring and Maintenance**

* **Apache Ambari:** Provides a centralized platform for provisioning, managing, and monitoring Hadoop clusters, offering a web-based dashboard to track the health and status of various components.
* **Apache Oozie:** Schedules and manages Hadoop jobs, automating the scheduling of ETL jobs and ensuring reliable job execution.
* **ELK Stack (Elasticsearch, Logstash, Kibana):** Collects and visualizes logs from Hadoop components, helping to monitor the system and troubleshoot issues efficiently.
* **Apache Zookeeper:** Maintains configuration information, naming, providing distributed synchronization, and group services, managing distributed applications by offering high-performance coordination services.
* **Apache Falcon:** Used for data management and policy enforcement in Hadoop clusters, providing data replication, lineage tracking, and notifications for critical alerts, ensuring data governance and reliability.
* **Regular Maintenance:** Includes tasks such as backups, performance tuning, applying security updates, and monitoring health using Ambari alerts and custom scripts.

**Conclusion** In summary, my role involves designing and managing comprehensive data pipelines that integrate various big data tools to ingest, process, and analyze large volumes of data. By leveraging the Hadoop ecosystem tools like Apache NiFi, Kafka, Spark, HDFS, and a suite of monitoring and orchestration tools, we have built a robust system that supports both batch and real-time data processing. This system enables us to deliver valuable insights to the business, driving data-driven decision-making and improving customer engagement.