Real-time data streaming service

Overview

Corporate business is producing a variety of big data in the field of innovative technologies such as IoT, autonomous driving, and AI. In order to make valuable use of this big data, data pipeline services to transmit, process, and analyze data in real time are essential.

This document describes the architecture that can efficiently respond to large data processing by utilizing both data batch processing and real-time streaming processing.

SDS Cloud provides various data streaming services, including **Kafka**, Spark, Hadoop, **Elasticsearch**, **Object Storage**, and **DB service** (RDB and NoSQL DB), for real-time processing and utilization.

Architecture Diagram

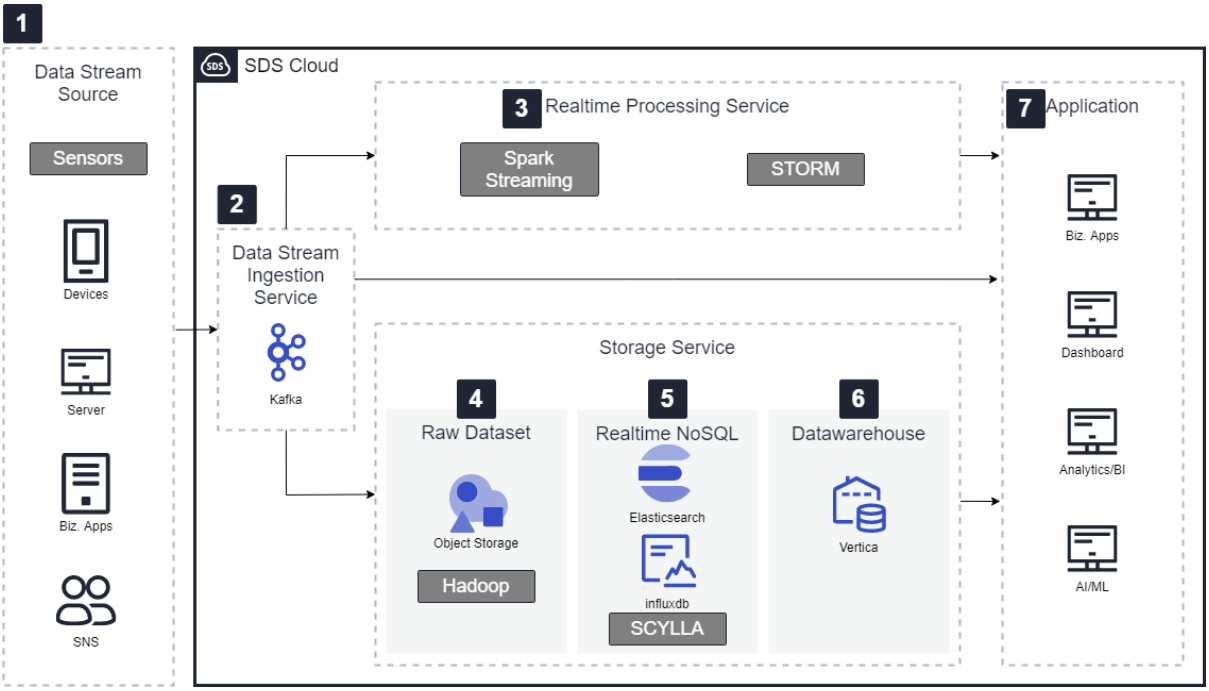

Figure 1. Real-time data streaming service structure using Kafka

1. Data Stream Source

Data streams come from a range of different sources, including mobile devices and other IoT devices, sensors of manufacturing facilities, SNS applications, and logs of system servers.

2. Data Stream Ingestion
   In order to process a large amount of data streams in real time, proper performance (in terms of latency and throughput), stability, and scalability are required.
   Kafka is an open source messaging solution with excellent scalability and stability, and is suitable for processing large-scale data streams. **Kafka's** various connectors enable easy connection to diverse data sources and using the data collected from one data source to be used in other services.

3. Real-time Processing
   Business requirements sometimes requires real-time processing by checking the data value in the streaming state. For this purpose, in-memory data processing open source software Spark Streaming and Storm can be used.

4. Raw Dataset Storage
   Object Storage and Hadoop (HDFS) services are available for stable storage that can be scaled out at low cost. These services are designed to prevent data loss and provide high durability. Data storage space can be expanded from GB to PB according to the size of the business, enhancing business agility and flexibility.

5. Real-time NoSQL
   You can collect and store (index) logs and metrics in various formats using Elasticsearch. In addition, many search functions (full-text search, Korean morphological analysis) are provided to quickly satisfy difficult search conditions. Data can also be analyzed and visualized in real time through Kibana.
   For storage, you can use InfluxDB, a DB specialized for processing time series data, or ScyllaDB, a high-performance data processing NoSQL compatible with Cassandra.

6. Data Warehouse (DW)
   Vertica is provided for analyzing large amounts of structured data in a cloud environment. As a native column store, the commercial solution has higher performance compared to other DW DBs and enables ML performance in the database without data migration based on the in-DB ML function.

7. Application
   Solutions that utilize real-time data streams or processed batch data include real-time dashboards, monitoring, analytics, BI tools, and AI/ML solutions.

Use Cases

A. Real-time anomaly monitoring for manufacturing facility management

Various sensor data is generated in real time in manufacturing facilities. It is important to constantly collect and analyze data in real time to ensure that the information needed for process control is accessible anytime, anywhere. Data generated by various sensors installed in manufacturing plant facilities is collected and classified by Kafka. This data can be processed by Spark Streaming to send anomaly notifications to users in real time. Raw data is stored in Elasticsearch to enable searching for specific conditions, and analysis based on data visualization.

B. Real-time fraud detection

The financial sector is required to conduct preliminary identification of the consumer's financial transaction patterns and in case of a sign of abnormality, request additional verification or provide notifications to the customer. For this purpose, you can easily build a fraud detection system with SDS Cloud's real-time Data Streaming services.

Pre-requisites

Customer License(BYOL) is required to build Vertica services.

Limitations

None

Considerations

None

Related Products

- Object Storage
- Elasticsearch
- Kafka
- Vertica
- Cloud Hadoop (To be available in 2022)
- Timeseries DB (To be launched)

Related Documents

- Server load balancing using GSLB