Data Infrastructure for Streaming Data Analytics

Stream Processors



Pre-written code to process streaming images. It assumed that the exactly-one semantic is applied and the image type can be obtained here.



Streaming Data Source



a) Images uploaded by customers



b) Stream of Image data from Kafka Stream (CustomerID reflected as NULL)



Data Ingestion, Transformation And Analytics



Cloud Pub: Ingest and distribute data reliably



Cloud Dataflow: Fast, simple, correct computations



BigQuery: Store data in data warehouse and data transformation



Data Storage and Archival Management



Cloud BigTable: Use of Garbage collection to remove data more than 7 days old to keep cost low



Real-time dashboards (BI)



Google Analytics



- a) Number of Images
- b) Type of Images
- c) Number and types of images by different customer groups

Requirements and Streaming Data Cloud Services Comparison

Concepts	Alibaba Cloud	AWS	AZURE	Google Cloud	IBM Cloud
Data Warehouse	MaxCompute	Athena, Redshift	Azure SQL	BigQuery	IBM Db2 Warehouse
Data Retention	Default - 24 hours	Default – 24 hours, 1-7 days (maximum 7 days)	-	-	-
SDK Support	MaxCompute Tunnel SDK	AWS SDK supports Android, Java, Go, .NET	Management .Net SDK	Apache Beam SDK	Eclipse SDK
Configuration	Writer plug-in	Days/Shards	-	-	-
Real-time Store	ApsaraDB	Amazon DynamoDB	Azure CosmosDB	Cloud Bigtable	IBM Cloud Object Storage
Cost	Pay-As-You- Go	Pay and use	Pay-As-You- Go	Based on the actual use of Dataflow batch or streaming workers	Based on instance per hour

- Able to support data from web application and Kafka Stream (based on Java and Scala)
- Able to set data retention to be more than 7 days (AWS does not meet the requirement)
- Source: https://wire19.com/streaming-dataservices-comparison-alibaba-cloudaws-azure-google-cloud-ibm-cloud/

References

- https://www.youtube.com/watch?v=DFtaavorJXY
- https://www.upsolver.com/blog/streaming-data-architecture-key-components
- https://wire19.com/streaming-data-services-comparison-alibaba-cloud-aws-azure-google-cloudibm-cloud/
- https://analyticsindiamag.com/top-7-data-streaming-tools-for-real-time-analytics/
- https://www.techradar.com/sg/best/best-cloud-analytics