Industry Examples of Distributed Systems

Vishwa Mohan

INTRODUCTION

- To implement the distributed functionality in our starGrad application, we will need tools to help us in the following areas:
  - Storage
  - Computation
  - Messaging
  - Database Management System
- These distributed tools do not have to be created from scratch. There are companies like Amazon, Apache, etc. that provide these functionalities as services.
- Let’s take a look at some of these Industry examples in detail.

STORAGE: AMAZON S3

- Amazon S3 is Amazon’s cloud solution for distributed storage.
- Data is logically grouped in buckets.
- It has replication/redundancy support to provide fault tolerance.
- Theoretically, there is no upper limit to the amount of data being stored.

AMAZON S3

- Data is distributed across multiple machines based on consistent hashing.
- This is an **eventually consistent** system.
- Programming language neutral
- It is available to end users as APIs over HTTP.

WHEN TO USE S3?

- When we have to store structured/unstructured data of very high volume/velocity.
- This data is not to be used for live applications but for analytical purposes
- For example:
  - Amazon tracks all of it’s users click data on its e-commerce website and dumps that into S3. From S3, all this data is dumped into the ML/Analytical pipeline for customising the user experience on Amazon.com.
  - Another use case of using S3 is that logs files of every service being provided by the AWS can be stored in S3

ALTERNATIVES TO S3

IBM Cloud
Object Storage

Microsoft Azure
Blob Storage

10
01

Google Cloud Storage

COMPUTATION: APACHE SPARK

- Apache Spark is an open-source, distributed cluster-computing framework.
- It is a solution for distributed computing
- It embodies the fault tolerance of distributed systems and offers lightning fast speed owing to its in-memory computation model and other optimization techniques.
- It follows a master-slave architecture, with a master node along with a cluster manager acting as a master and several worker nodes acting as slaves.

WHEN TO USE SPARK?

- Batch data processing
- Real-time data processing
- Complex ML computations and complicated graph algorithms
- For example, NASA leverages Spark for log data processing and advanced analytics.
- PySpark, which is a Python API for Spark, is widely used for Data Science and Machine learning when the data sets are simply too large for a single computer.

ALTERNATIVES TO SPARK

Flink

hadoop
Map Reduce

amazon
KINESIS

presto

MESSAGING: APACHE KAFKA

- Kafka is a distributed messaging/streaming system.
- Fault-tolerant, highly available system
- Provides message ordering guarantee over a distributed design - reliable
- Connects distributed producers and consumers
- Horizontally scalable - Practically no upper limit to the amount (volume/velocity) of messages served, attributed to its distributed foundation

WHEN TO USE KAFKA?

- Building non-blocking systems (Blocking can be seen as making a system unavailable)
- Asynchronous systems
- Streaming data ingestion systems where data is collected in real-time
- For example, all the click events on LinkedIn are first moved to Kafka, which acts as the ingestion gateway for the big data ecosystem
- Netflix uses Kafka for its real-time monitoring and event processing pipelines

ALTERNATIVES TO KAFKA

APACHE ACTIVEMQ™

RabbitMQ

DBMS: MongoDB

- It is a distributed NoSQL database management system
- It satisfies transactional (OLTP) needs.
- It can support analytical (OLAP) requirements as well owing to its distributed nature.
- It has the advantages of both SQL and NoSQL databases.
- It offers very fast read and write speeds, which can be scaled easily as well.

WHEN TO USE MONGODB?

- When we need to deal with structured data, but the volume of data that needs to be stored can’t fit in a single machine. (MongoDB uses JSON-like documents with **optional schemas**. This can be considered semi-structured).
- Designed specifically for high load websites.
- Example: Adobe relies on MongoDB data management for its cloud data, which is in the scale of petabytes.

ALTERNATIVES TO MongoDB

Thank You