Chapter Six

Big Data Processing Concepts

Parallel Processing

A task can be divided into three sub-tasks that are executed in parallel on three different processors within the same machine.
Distributed Processing

Hadoop

a versatile framework that provides both processing and storage capabilities.

Processing Workloads in Big Data

A batch workload can include grouped read/writes to INSERT, SELECT, UPDATE and DELETE.
Transactional workloads have few joins and lower latency responses than batch workloads.

Utilize it to support batch processing of bulk data and real-time processing of streaming data.

The symbol used to represent a processing engine.
An illustration of a MapReduce job with the map stage highlighted.

Map Reduce - Map Stage

The combine stage groups the output from the map stage.
A summary of the combine stage.

Map Reduce -- Partition Stage (optional)

The partition stage assigns the output from the map task to reducers.

A summary of the partition stage.
During the shuffle and sort stage, data is copied across the network to the reducer nodes and sorted by key.

The reduce stage is the last stage of the reduce task.
A summary of the reduce stage.

Map Reduce Example

An example of MapReduce in action.

Understanding Map Reduce – Tsk Parallelism

A task is split into two sub-tasks, Sub-task A and Sub-task B, which are then run on two different nodes on the same dataset.
A dataset is divided into two sub-datasets, Sub-dataset A and Sub-dataset B, which are then processed on two different nodes using the same function.

An instance where three lines constitute a single record.

Key considerations when developing a MapReduce algorithm

- Use of relatively simplistic algorithmic logic, such that the required result can be obtained by applying the same logic to different portions of a dataset in parallel and then aggregating the results in some manner.
- Availability of the dataset in a distributed manner partitioned across a cluster so that multiple map functions can process different subsets of a dataset in parallel.
- Understanding of the data structure within the dataset so that a meaningful data unit (a single record) can be chosen.
- Dividing algorithmic logic into map and reduce functions so that the logic in the map function is not dependent on the complete dataset, since only data within a single split is available.
- Emitting the correct key from the map function along with all the required data as value because the reduce function’s logic can only process those values that were emitted as part of the key-value pairs from the map function.
- Emitting the correct key from the reduce function along with the required data as value because the output from each reduce function becomes the final output of the MapReduce algorithm.
Processing in Realtime Mode

- Data processed in memory
- aka stream processing
- Interactive mode – query processing in realtime
- Speed, Consistency and Volume (SCV) principle.

Venn diagram summarizing the SCV principle.

Realtime Data Processing Example in a Big Data environment.
Realtime Bigdata Processing and SCV

Realtime Bigdata Processing and SCV

- **Realtime** Bigdata Processing and SCV

  - While designing a real-time Bigdata processing system, the SCV provide means to keep in mind. In light of this, it is quite certain that in the near future, the focus will be on real-time Bigdata solutions. It is important to note that in the past, Bigdata solutions were often designed using batch-oriented solutions. However, the need for real-time processing is becoming more critical in today's fast-paced environment. Realtime Bigdata processing solutions are more challenging to design and implement than their batch-oriented counterparts.

  - In the context of real-time Bigdata processing, the IT team believes that the primary challenge is the need to process data in real-time. This means that the system must be able to handle incoming data streams and produce results in real-time. The system must be able to handle large volumes of data and provide insights in a timely manner. The goal is to provide real-time analytics to support decision-making.

Complex Event Processing

- **Complex** Event Processing

  - During CEP, a number of real-time events often coming from disparate sources and arriving at different time intervals are analyzed simultaneously for the detection of patterns and initiation of action. Rule-based algorithms and statistical techniques are applied, taking into account business logic and process context to discover cross-cutting complex event patterns.

  - CEP focuses more on complexity, providing rich analytics. However, as a result, speed of execution may be adversely affected. In general, CEP is considered to be a superset of ESP and often the output of ESP results in the generation of synthetic events that can be fed into CEP.

Realtime Bigdata processing and SCV and MapReduce

- **Realtime** Bigdata processing and SCV and MapReduce

  - The IT team believes that the processing in real-time and batch processing models are quite similar. Instead of getting a list of events, the system processes data in real-time, and the output is a list of results. The system then processes data in real-time and the output is a list of results. The system then produces an output in real-time.

  - The IT team takes the following approach. In the reducer, the input is a list of events, and the output is a list of results. The system then produces an output in real-time.

Realtime Bigdata processing and SCV and MapReduce

- **Realtime** Bigdata processing and SCV and MapReduce

  - The IT team believes that the processing in real-time and batch processing models are quite similar. Instead of getting a list of events, the system processes data in real-time, and the output is a list of results. The system then processes data in real-time and the output is a list of results. The system then produces an output in real-time.

  - The IT team takes the following approach. In the reducer, the input is a list of events, and the output is a list of results. The system then produces an output in real-time.

Realtime Bigdata processing and SCV and MapReduce

- **Realtime** Bigdata processing and SCV and MapReduce

  - The IT team believes that the processing in real-time and batch processing models are quite similar. Instead of getting a list of events, the system processes data in real-time, and the output is a list of results. The system then processes data in real-time and the output is a list of results. The system then produces an output in real-time.

  - The IT team takes the following approach. In the reducer, the input is a list of events, and the output is a list of results. The system then produces an output in real-time.

Realtime Bigdata processing and SCV and MapReduce

- **Realtime** Bigdata processing and SCV and MapReduce

  - The IT team believes that the processing in real-time and batch processing models are quite similar. Instead of getting a list of events, the system processes data in real-time, and the output is a list of results. The system then processes data in real-time and the output is a list of results. The system then produces an output in real-time.

  - The IT team takes the following approach. In the reducer, the input is a list of events, and the output is a list of results. The system then produces an output in real-time.

Case Study

- **Case Study**

  - The IT team believes that the processing in real-time and batch processing models are quite similar. Instead of getting a list of events, the system processes data in real-time, and the output is a list of results. The system then processes data in real-time and the output is a list of results. The system then produces an output in real-time.

  - The IT team takes the following approach. In the reducer, the input is a list of events, and the output is a list of results. The system then produces an output in real-time.