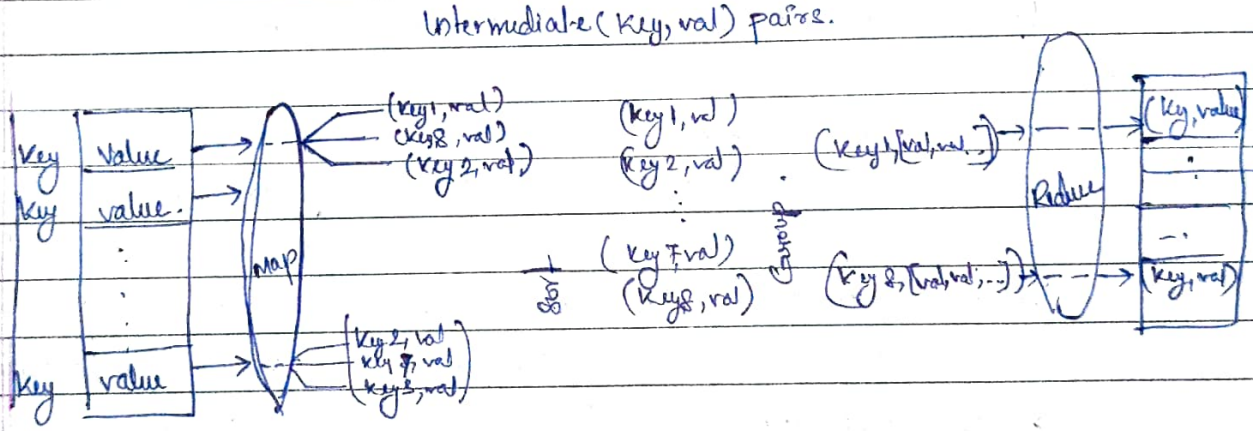
10/08/21

CS 468-A
CLOUD COMPUTING.

Christy Varghese
CSE A
Roll No. 34.

① MapReduce

- A software framework which supports parallel and distributed computing on large data sets.
- MapReduce s/w framework provides an abstraction layer with the dataflow and flows of controls to users, & hides the implementation of all data flows steps such as data mapping, partitioning, synchronization, and scheduling.



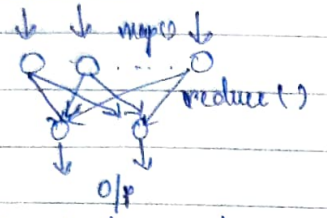
The Map func. is applied in parallel to every i/p pair, and produces new set of intermediate (key, value) pairs as follows:

$$(key_1, val_1) \xrightarrow[\text{function}]{\text{map}} \text{List}(key_2, val_2)$$

Then the MapReduce library collects all the produced intermediate pairs from all i/p pairs & sort them based on the 'key' part. Finally, the Reduce func. is applied in parallel to each grp producing the collection of values as o/p as illustrated here

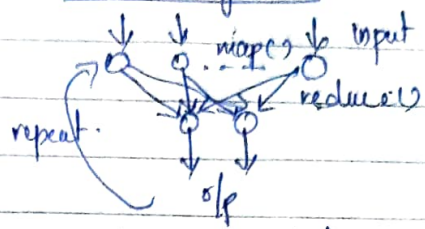
$$(key_2, \text{List}(val_2)) \xrightarrow[\text{func}]{\text{Reduce}} \text{List}(val_2).$$

(2) Map Reduce



- Distributed search
- Calculation of pairwise distances for sequences (BLAST)
- Distributed sort.
- High-energy physics (HEP) histograms.
- Information retrieval.

Iterative Map Reduce.



- Data mining including
  - clustering
  - K-means
  - Deterministic annealing clustering
  - MDS.
- Linear algebra.
- Expectations max^m algorithms.

(3) BigTable comprises of : a client library, a master server that coordinates activity, and many tablet servers. Tablet servers ~~Each~~ ~~tabl~~ The master assign tablets to tablet servers & balances tablet server load. Each tablet server manages a set of tablet. It handles read/write requests to the tablets it manages and splits tablets when a tablet gets too large. Client data does not move through the master clients communicate directly with tablet servers for read/write. The internal file format for storing data is Google's SSTable which is persistent, ordered, immutable map from keys to value.

(4) Data Governance framework

The major 5 components of data governance framework are :

Phase 1 : Understanding the Problem.

Inorder to maximize the efficiency, it is important to talk to the executives and managers across the org. and ~~st~~ understanding what main points they'd like to solve.

Phase 2: Strategizing and Planning

During this phase, we outline both strategic and tactical approaches focussed on ~~tacti~~ achieving the long-term goals.

Phase 3: Organizing

After discussing and choosing a correct plan, we also need to gather all the information together so we will require a data governance council to organize all the information.

Phase 4: Communicating

Data Governance programs extend over a period of time and program fatigue can be common. Good communications lead to continued buy in from the program sponsors.

Phase 5: Executing

After gathering all the data, organizing it and developing plan, the final phase is to execute the plan.

(5)

(b) Pig Latin

→ It is a high-level data flow language developed by Yahoo that has been on top of Hadoop in the Apache Pig project.

→ Apache Pig is a high-level platform for creating programs that run on Apache Hadoop.

→ Pig can execute its Hadoop jobs in MapReduce, Apache Tez or Apache Spark.

→ It abstracts the programming from the MapReduce idioms into a notation which makes MapReduce prog. high level, similar to that of SQL for relational database management systems.

→ It can be extended using user-defined functions (UDFs) which the user can write in Java, Python, JS, Ruby and then call directly from the language.

6 (a) Secure Software Development Life Cycle (SecSDLC).

- The SecSDLC involves identifying specific threats and the risks they represent, followed by design and implementation of specific controls to counter those threats and assists in managing the risks they pose to the org.
- The SecSDLC must provide : consistency, repeatibility & conformance.

→ Phase 1 : Investigation : Define project processes and goals, and document them in the pgm security policy.

→ Phase 2 : Analysis : Analyze existing security policies and programs, analyze current threats & controls, examine legal issues. & perform risk analysis.

→ Phase 3 : Logical design : Develop a security blueprint., plan incident response actions, plan business responses to disaster and determine the feasibility of continuing and outsourcing the project.

→ Phase 4 : Physical Design : Select technologies to support the security blueprint, develop a definition of a successful soln, design physical security measure to support tech. solutions.

Phase 5: Implementation / ~~Maintenance~~ : Buy or develop security solns. At the end of this phase, present a tested package to management for approval.

Phase 6: Maintenance : constantly monitor, test, modify, update and repair to changing threats.

6(b)(i) Demand-Driven Resource Processing

→ This method adds or removes computing instances based on the current utilization level of the allocated resources.

→ When a resource has surpassed a threshold for a certain amount of time, the resource could be decreased accordingly.

→ In general, when a resource has surpassed a threshold for a certain amount of time, the scheme increase that resource based on demand.

→ The scheme does not work out right if the workload changes abruptly.

6(b)(ii) Event-Driven Resource Processing

→ This scheme also adds or remove machine ~~be~~ instances based on a specific time event.

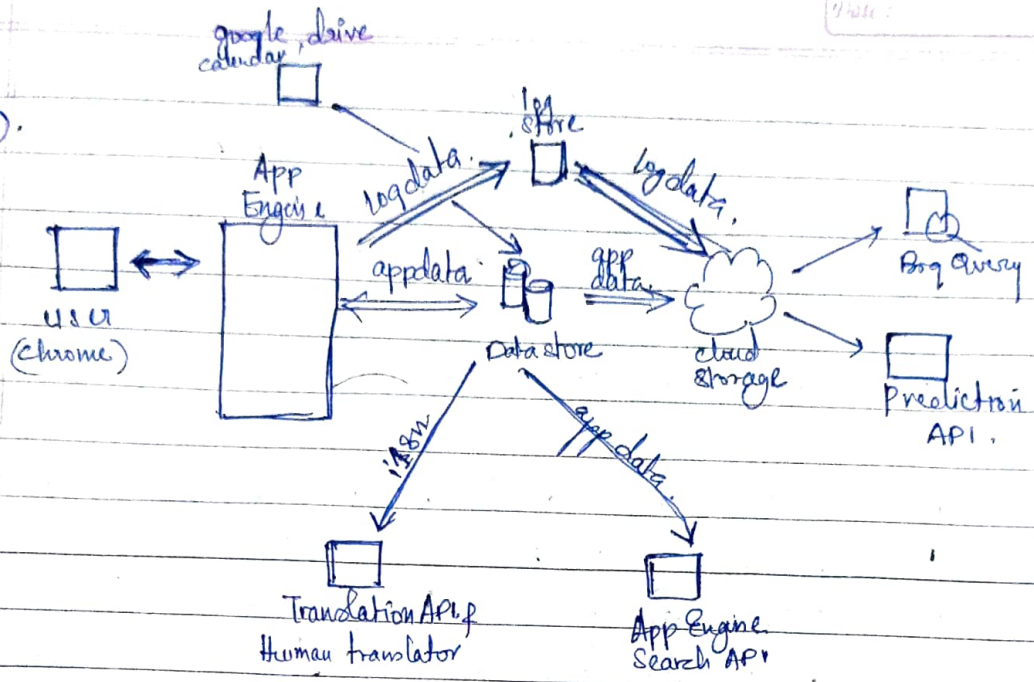
→ During these events, the number of users grows before the event period and then decreases during the event period.

→ This scheme anticipates peak traffic before it happens.

→ This method results in a minimal loss of QoS, if the event is predicted correctly.

→ Otherwise, wasted resources are even greater due to events that do not follow a fixed pattern.

5 (a).



7 (b) Seven security issues on cloud vendor

(a) Data segregation: Data in the cloud is typically in a shared env. alongside data from other customers. Encryption is effective but isn't a cure all.

(b) Recovery: Even if you don't know where your data is, a cloud provider should tell you what will happen to your data & service in case of a disaster.

(c) Long-term viability: Ideally, ~~your~~ cloud computing provider will never go broke or get acquired and swallowed up by a larger community.

(d) Privileged user access: Sensitive data processed outside the enterprise brings with it an inherent level of risk.

(e) Regulatory compliance: Customers are ultimately responsible for the security and integrity of their own data.

(f) Data location: when you use the cloud, you probably won't know exactly where your data is hosted. In fact you

not even know what country it will be stored in.

(g) Investigative support : Investigating inappropriate or illegal activity may be impossible in cloud computing.

(3)(a) Cloud computing service provides a remote service to its users known as cloud service. These scalable solutions are managed by a third party and provides user with access to computing services such as analytics or networking via the internet. They offer powerful benefits for the enterprise, from greater productivity and enhanced efficiency to significant cost reductions and simplified IT management.

Google calendar is the most used calendar today. In all the android device has this application by default. For me, google calendar helps me to organise personal meetings, public holidays, birthdays of my friends and relatives and also helps to track me on the events. Google calendar can be collaborated with Email, Schedules, Contact lists etc..