Page No.
Date : / /

Building Web Scale Application

* Web Crawling Process:

=> The web crawling Process involves systematically browsing the web to index and retrive information from the website.

The Process begins such a set of seed URL's which are the initial web pages that the crawler visits.

The crawler send HTTP requests to the seed URLs and fetches the corresponding web pages.

Once the web pages are fetched, the crawler parses the HTML content to extract links to other web page.

The extracted URLs are filtered based on predefined rules.

The content of the fetched page is stored and indexed for later use.

SMVS
Page No.
Date : / /

The extracted URLs are recursively followed by crawler and fetching more web pages.

Crawlers implement rules to avoid ~~ov~~ overwhelming web servers and write into the robots.txt File.

The goal of web crawling is to efficiently gather relevant data while respecting the constraints.

-> Advantages :

1 Automated Data Collection : It automate the process of gathering large volumes of the data from multiple website.

2 Scalability : Web crawlers can scale to cover a vast number of web pages across the internet.

3 Data For Analysis : Crawlers provides raw data that can be used for various purposes.

4 Efficient Indexing : Crawlers help search engine to index website.

SMVS
Page No.
Date : / /

* Web Graph Mining:

=> Web Graph mining involves analyzing the web as a directed graph.

In Web Graph, Web pages are vertices and hyperlinks between them are edges.

Two prominent algorithms used in the web graph mining for Ranking web page.

1) Page Rank
2) HITS

1 Page Rank:

Page Rank was developed by Google's founders to rank web pages based on their link structure.

It assign a rank or score to each web page, indicating its importance.

This Algorithm links from one page to another with links and High-ranked pages being more valuable.

SMVS
Page No.
Date : / /

-> Formula:

$$PR(i) = (1-d) + d \sum_j \frac{PR(j)}{L(j)}$$

where, PR(j) -> Page Rank of Page j linking to Page i
L(j) -> No. of outgoing links from page j
d -> Damping Factor

PageRank assigns a Global Importance score to web pages.

2 HITS Algorithm:

HITS stands for Hyperlink-Induced Topic Search which is another algorithm for ranking web pages.

It is used the concept of hubs and authorities.

Authorities: Pages that considered authoritative because they are linked to by many hubs.

Hubs: Pages that link to many authoritative pages on topic.

Date : / /

Each web page has two score:
An Authority Score, A Hub Score.

Authority Update $A(p) = \sum_{q \in H} H(q)$

Hub Update $H(p) = \sum_{q \in A} A(q)$

=> Web Graph Mining Advantages:

1 Improved Search Quality

2 Scalability

3 Authority and Trust Evaluation

4 Efficient Web Crawling

5 Link Prediction

* Distributed System:

=> A Distributed system is software that coordinates the actions of multiple computers.

It is enabling them to work

SMVS
Page No.
Date : / /

together toward a common goal.

This coordination is achieved through message exchanges between the computer.

Distributed system are fundational to mordern computing, including data centers, cloud computing etc.

Computers in a distributed system do not share storage and processing power.

This architecture is scalable as adding more computers to the system can increase capacity without creating resource contention.

Local Area Network (LAN) are commanly used in environments like data centers to connect a large number of servers.

Peer-to-Peer network are special type of overlay networks which built on top of the physical network.

SMVS
Page No.
Date : / /

=> Property For Distributed System

1 Reliability:

It ensure that user action are not cancled due to the hardware or software failures.

2 Scalability:

It ensure that, Distributed system has a capacity to handle more work without slowing down.

3 Availability:

It ensure, that Distributed system has ability to keep services running smoothly even when some servers fail.

4 Efficiency:

It is focusing on how well the system performs its task or how quickly the system provides the first result.

SMVS
Page No.
Date : / /

* Failure Management:

=> In a Distributed System, failure management is critical due to the complexity of network.

In centralized system, failure management is simple compared to distributed system.

-> Principles of Failure Management:

1 Independence:

Each node or server should work independently, so if one fails it can be fixed without affecting others

This reduces the need for coordination and keeps the system simpler.

2 Replication:

Replication mean having backup nodes or server that can take over if one fails.

This ensures the system continues

working smoothly with minimal downtime without affecting by Failure.

=> Failure Management in Centralized Systems:

In Centralized system, all resources are managned by a single machine or server.

Managing Failures in Centralized system is simpler because there are only one machine to monitor and recover.

-> Recovery Mechanisms:

- Logging: The system keeps a log of actions or transaction in a file.

  If a Failure occurs, the log helps restore the system.

- Backups: Regular Backups of the system data allow recovery in case of Failure.

=> Failure Management in Distributed System:

Distributed systems consist of multiple nodes or servers working together.

Failure management in such system is more complex because failure can happen at any nodes or server.

-> Recovery Mechanisms:

- Replication: Data and services are replicated across multiple nodes so, that if one fails another can take over.
- Failover: When a node or server fails, its task are transfered to another node to maintain system functionality.
- Consensus Protocol: It use algorithm like Two-Phase Commit to ensure that all nodes agree on transaction.

If a failure occurs during a transaction, the system can either complete or roll back the changes.