NAME:- PREKSHA PATEL
SAPID:- 60004210126
BRANCH:- COMPUTER ENGINEERING
DIV:- C2; BATCH:- 1

BIG DATA INFRASTRUCTURE (BDI)

EXPERIMENT NO.-1

Case Study on LinkedIn – Leveraging Big Data Analytics

Big Data refers to processing huge volumes of data that are beyond traditional processing of RDBMS. Data in today's digital and tech world is highly diverse in variety, type, source, velocity, veracity etc. which needs real-time handling, pre-processing and summary analysis for accurate pattern recognition, association, correlation, regression, visualization, fraud detection and effective decisions. Today we are all surrounded with structured, semi-structured and unstructured data like operational day to day data, log files, CCTV footage, audio and video streaming, multimedia, chip and circuit generated data, sensors etc which are not only complex to handle but also to dig and mine out productive results. Organizations do intense refining, pre-processing, filtering and mining on big data for machine learning projects and other analytical applications.

Context to LinkedIn

LinkedIn is the world's largest networking community for professionals in social media with more than 500 million profiles across 200+ countries. People prefer to create their LinkedIn profiles not only for showcasing their professional skills and achievements but majorly to connect with established corporate leaders and mangers for better insights, job opportunities, corporate news etc. LinkedIn is an apt platform where people can share their expertise and connect with similar domain and like-minded professionals for discussion and updates on various issues of interest, as it provides a semi-formal and traditional environment. People on LinkedIn vary from job-hunters, freshers to top professionals and each one of these categories avail quick service on LinkedIn by predicting what kind of information is needed, when and of which nature.

At LinkedIn, big data is more about business than data. LinkedIn heavily relies on Big Data Analytics for managing all profiles and their security and privacy along with providing relevant information to authorized users. LinkedIn is a large platform of social network not just in terms of revenue or members but also in terms of its multiple data products. LinkedIn processes thousands of events every day and does tracking of each user activity for better results and search queries. Big Data here plays a vital role in various interactions in social graph and so a number of big data designers, engineers, scientists and analysts are not only deployed but also networked. LinkedIn uses a lot of data for its recommender engine to create data products and build a comprehensive picture of a profile. Data scientists and analysts use big data to derive valuable business insights and performance metrics that lead to profitable decision making for sales, marketing and other functional areas. LinkedIn knows where you should apply for a job, whom you should connect with and how your skills stack up against your competitors and colleagues as you look for your dream job.

Application of Apache Hadoop in LinkedIn for leveraging Big Data technologies

Following are various big data techniques and tools deployed in LinkedIn:

1. Giraph-enables computations and interactions on social graph

2. Avatar- enables search of ‘who viewed My profile’ feature using OLAP3. Voldemort- provides NOSQL interface for distributed storage and processing
4. White Elephant-provides visualization dashboards giving every event summary
5. Kafka- used for tracking all LinkedIn events like page and profile views, updates, searches etc.
6. Azkaban-used for managing workflow system and scheduling jobs

Data products used at LinkedIn

1. Skill endorsements- used by recruiters for extracting skills and expertise based right candidates. It is an information extraction data product of LinkedIn that uses big data technology.
2. News Feed Updates- provides relevant data through data analytics and machine learning algorithms of Hadoop for instant prototyping and new update testing.
3. People you may know-It is the best feature of LinkedIn to enable connections. Earlier this feature used Python code but now Hadoop batch processing to filter both online and offline data.
4. Join You May Be iNterested In- for looking top talent, this feature provides searchable job titles, skill sets and connections. It is used by 90% of Fortune 100 companies for hiring top talent and 89% of professionals to land a job. This feature provides 50% of website engagement.

Big Data Infrastructure used at LinkedIn

1. Apache Hadoop
2. Hadoop Distributed File System (HDFS)
3. Pig
4. Hive
5. Zookeeper
6. Azkaban
7. Kafka
8. Voldemort etc.

Conclusion

LinkedIn is one of the pioneer in professional networking and is heavily using technology, big data, machine learning and artificial intelligence for effective and accurate social profiling, connects and networking. Apache Hadoop powers commonly used features both on the website and mobile app. It stores, manages and analyses volumes of data from diverse sources stored in data warehouses and marts, providing distributed computations in real time and descriptive statistics for live and interactive dashboards and thus enabling ad-hoc analysis and other searches effective and result driven.