CSE1902 Industrial Internship

(Online Course)

Day to Day Activities Diary Report
Duration: 33 days
Fall Semester 2021-22

by
Sharadindu Adhikari
19BCE2105

Sl.	Date	Activity
1.	27 June	Started Module 1 of Futureskill Nasscom Digital Vidya's 'Introduction to Big Data' course. Learned about Evolution of Data from ERP and CRM, and how things have come to develop into Big Data as we know it today. Couple of interesting facts, one of them being the Four Vs. The 5th V. And the Problems with Traditional Large-Scale systems and Grid Computing: how in a storage system, nodes are divided (acting as bottlenecks); and how they are distributed.
2.	28 June	Day 2. Studied about overcoming the problems with traditional system of grid computing. The need for multi-core distribution and how fallacies of Moore's law is limiting CPU speeds. Learned more about distributed parallel processing: how a group of independent and geographically dispersed computer systems take part to solve a complex problem. And a couple of its characteristics, briefly.
3.	29 June	Started with big data grid computing. And its solution landscape. Read through some of the industry insights, governance compliances for financial services, EDW optimisation and risk management. Also watched its corresponding lecture about why banking sector is aggressively adopting big data tech. Followed by [big data's hand in] Healthcare Data Lake: IOT, Electronic Health Records, Reduction of Fraud-waste-abuse. Examined some key use cases: customers' segmentation, sentiment analysis, call centre analysis and fraud detection. And the key IT consideration before implementing any big data solution to such problems.
4.	30 June	Firstly, read about redundant physical and security infrastructures. And then different operational databases. A couple of points regarding organisation of data services and tools. Followed by the names of big data analytic providers and solutions: current offerings, strategies, market preferences, etc. Thereafter I watched the designated video which introduced me to Big data as a career — different job titles: Data scientists, ML professionals, Business analysts, Data engineers, Solution architect, Consultants, etc. Thereafter I analysed a couple of bar graphs regarding adoption of big data tools in different industries.

		Followed by the details of MNCs, Consultant groups, and International Banks who are adopting the big data technology.
5.	1 July	Day 5. I'd started the 3rd quarter of 2021 with reading the business opportunities, maturity index, data transformations, and analysing different models of the big data industry. In the big data maturity index model, I've learned about business insights, optimisation techniques, data monetisation, and business metamorphosis. Watched the video and studies some trends, benchmarks, and how to launch a prototype of pilot project. Followed by evaluating business insights, identifying supporting business decisions, studied how to determine the analytics, data enrichment and data transformation. Analysed how to find the target customer, the importance of investing time in research and understanding potential ecosystem players, at the same time focusing on the product development.
6.	2 July	Started studying the CPG Domain and other Business opportunities: primarily focusing on demand-based forecasting, supply chain optimisations, trade promotion effectiveness and markdown management. Also gave much heed to market basket analysis, price-yield optimisations and customer loyalty programs. Thereafter, I watched the video and learned about different IT stake holders — how they define data strategy, incorporate insights into apps, build analytic models, and implement big data architecture. At the day's end, I've briefly went through the Solution Engineering division to understand how organisations make money through sales and marketing, operations and finance. And identified some key business initiatives: how business stakeholders and their roles are responsible for key performance indicators, critical success factors and desired outcomes. And how their expectations are directly correlated to the timeframe of delivery.
7.	3 July	The next day I thought what better way to start off than some barnstorming, which I did for big data business impacts. In it I learned how to mine the more detailed data at the lowest level of transaction granularity and how to integrate new unstructured data sources to enable more robust and complete decision, at the same time providing real-time low-latency data access. After that I studied how they generally break down the business initiatives into use cases. Like what business decisions the stakeholders are trying to make, how they are targeting different personas, and how they are using data and analytics algorithms to capture user experiences. After it is being dealt with, I'd to learn how to prove out these use cases. They're typically done by gathering required data, defining and executing data transformation, fine-tuning the analytics models and developing mockups. Followed by designing & implementing the bigdata solutions: instrumentation strategy, business intelligence, etc.
8.	4 July	With the start of my 2nd week on this course, I'd my focus on different customer behavioural analytics — interest-based engagement levels, customer loyalty and satisfaction, their journey stages and purchasing behaviours, etc. Thereafter I read about some off-the-grid factors which heavily affect these

		analytics. Watched the corresponding video immediately and found that, most of them focussed on predictive maintenance, marketing effectiveness, fraud detection, and network optimisation. With this, I'd started learning about big data impacts: different business values, data analytics, how things are unfolding nowadays and how the industry is shifting and changing their policies to accommodate for the future. Then I learned about the differences between a BI Analyst and a Data Scientist — how they gather, capture and assess data; how they cleanse the data for provisioning an analytical workspace; and how they plan and build different analytic models, assess their reliability and prepare reports.
9.	5 July	July 4th was the end of Module 1. Before I could move forward with the course, I'd to turn in an assignment. It asked of me to identify use cases from industries of my choice and elaborate on how big data analytics can be used to transform those businesses. For me it was quite interesting. I'd written about the 4th Industrial revolution and IoT; how earlier problems of data acquisition has vanquished; and how different operation improvements help support strategic decision making. And of course, some statistics to back up these points.
10.	6 July	Watched the next video in order, and with that I started Module 2. Didn't go into much detail; just briefly studied about data warehouses, their structure and types, and a couple of terminologies commonly used in big data, like in-memory analytics, in-database processing, symmetric multiprocessor system (SMP), and massively parallel processing (MPP). It was a reading-heavy portion, and after getting done with it, I called it a day.
11.	7 July	Day 2 of Module 2. Read about parallel systems, and how they differ and combine with distributed systems. Learned about CAP Theorem and how RDBMS, MongoDB, CouchDB, and Date Models help make it. Then studied briefly about the Hadoop ecosystem and different big data technology landscapes. After that I began the Hadoop installation on my PC, following all the instructions diligently. Studied about different HDFS Commands. There were so much stuff, it took me a while to read through them. Studied about HiveQL and how it enables users to perform tasks using MapReduce concept without explicitly writing the code in terms of the map and reduce functions. And how data stored in HDFS can be accessed through HiveQL, which contains the features of SQL but, runs on the MapReduce framework. And lastly, a couple of ways to accessing the Hive: Hive CLI, Web Interface, Server, and JDBC/ODBC.
12.	8 July	Started the day with learning the Hive Architecture. Followed by different data types (primitive and complex) and built-in functions; and how CLI, Metastore, Task Tracer, Drive (Query, Complier, Executor), etc. comes together in the Arch. Other data definition language functions (create, alter, drop, show, truncate, delete). Learned a couple of DDL commands as well (to show and drop database tables in DigitalVidya). As well as Select Statements. Next up was Pig, which is a scripting language for exploring large

		datasets. It was designed to be extensible and is made up of 2 pieces: the language used to express data flows called the Pig Latin, and the execution environment to run it. Thereafter I learned about data types in Pig (bytearray, chararray, tuple, bag, maps, bigdecimal, boolean, etc.) and called it a day
13.	9 July	Started just where I left off the previous day. Watched the corresponding video portion about Pig Queries (load, for_each, filter, and dump) and learned how to get on with Sqoop. Learned about how Sqoop let us create step by step procedures on how to import data from MySQL (as well as MariaDB, which is a sister branch) to HDFS. Then I read a brief discussion about Sqoop's practical aspects to creating and using databases, rows, and tables. After I'd dealt with this theoretical portion, I'd to watch a couple of tutorials to run MR Word Count in Hadoop. There were 7-8 steps and I'd followed them well enough to run past the question which would be asked next.
14.	10 July	The final part of Module 2 began with understanding the concepts of Map & Reduce, and how the terms combined have a solid meaning and use. There were several stages in the MP Word Count — input files, individual mapper, map key value splitting, sort & shuffle, reducing key value pairs, and final output. After that I watched the video portion about Hadoop namenode, reduce layer, and HDFS layer. And how TaskTracker and DataNodes are entangled with them. I learned that JobTracker knows everything about submitted jobs. It divides job into tasks and decides where to run each task. Thereafter, I read through data flows of Map Reduce; its phases of the job and separate graphs. Followed by key-value pair generation, where I learned that MapReduce framework operates exclusively on <key, value> pairs.
15.	11 July	The last day of studying Module 2. I'd read about YARN MR detail flow, watched its corresponding video, and learned how client nodes, resource manager nodes, node manager nodes and even HDFS intertwine together. Thereafter I learned about the concept of Mapper — different input formats, splits, record readers, keys, and collectors. Followed by the size details of Map Tasks and the concept of Reducer. Watched the last video portion about speculative execution of nodes and job schedulers. And Hadoop fault tolerance, which discussed some crucial points like what if a task or a data node fails, or what if a name node or job tracer fails. The final portion I'd read explained submission, initialisation, monitoring, and progress of Map Reduce job.
16.	12 July	Before I could move on to the next phase of the course, I'd to solve the 2nd Assignment. By any means it was tedious and challenging. It'd asked of me to perform a couple of tasks on Hadoop. I had to run map and reduce codes, perform data storage and retrieval operations, and do some batch processing operations. 9 steps and 4 hours later, I turned it in.
17.	13 July	Halfway through Week 3, I started with Module 3. Watched the introductory video and learned about data ingestion and its steps — extract, transform, load. I was instructed, in the career of a data scientist, to expect difficulties and to plan accordingly. That there are several misconceptions and hype over it. Many

		enterprises begin data analytics projects without understanding this, and then they become surprised or disappointed when the data ingestion process does not meet their initial schedules. After that, I learned about automating data ingestion, and the drawbacks of curating it manually.
18.	14 July	The next day I'd read about the use of AI in Big Data, and how some firms are helping channel it. The corresponding video portion discussed some more about system automation and explained a few examples of the processes. Thereafter, I learned about how to make it a self-service, govern the data to keep it clean (responsibility includes: defining Schema, cleansing rules, decision to ingest particular data into data source, treatment of dirty data), and advertising the cleansed data. It is very important that organizations should implement a publish-subscribe model with a registry of previously cleansed data available for lookup by all other users.
19.	15 July	On July 15th, I started up with the famous NoSQL Database: MongoDB. Watched the first 17 minutes of the corresponding video and read about the reasons to learn MongoDB, in detail. Learned the importance of aggregation framework, binary json format, sharding, ad hoc queries, schema-less, capped collections, indexing, replication, and MongoDB management services. More on it I read about how scaling causes major problems with any web/mobile application, and how MongoDB is helping tackle it. Finally I get to learn about indexes and how, to improve performance of searches, they are created.
20.	16 July	I started where I left off, with the exact video I was watching the previous day, and completed the last 25 minutes of it. In it I learned about the benefits of mongo database: load balancing, sharding, flexibility, and speed. It was discussed with some really good animations; really liked that part. After that, I read about the drawbacks and limitations of MongoDB. There weren't many, but data consumption is generally high due to de-normalisation, and there really aren't any default transaction support. The design in itself is inconsistent.
21.	17 July	Day 3 of learning MongoDB. I watched the corresponding video and read about data types, database commands, relational operators and operational commands. It was a long read. After that, I'd learned about the differences between relational database management systems and mongo database, primarily w.r.t differences in update and select queries. With as many as 10 unique points, I read the differences and called it a day.
22.	18 July	Started the day with watching the last video of MongoDB. It was more than an hour long, so took a while. To get in-hand experience, I'd to do an exercise first, before moving on to the next portion of the module. In it I'd to create a mongo database called "student", and implement a lot of SQL queries — like creating a specific collection, inserting a couple of documents with fields, display them, sort some of them, delete a couple, display the rest, and update it. This is what I did for the day.
23.	19 July	I'd started the final part of the Module with learning about data pipelines and how building it become a commonplace for

		everything related to data. I read that typically, the destination for a data pipeline is a data lake, such as Hadoop or parquet files on S3, or a relational database, such as Redshift. After that I started watching the corresponding video on big data pipelines, and covered the first 20 minutes of it, learning about different components: compute, storage, and messaging. Followed by reading about their deployment and finally exploring questions like who owns the data pipeline, which teams will be consuming data, and who will QA the pipeline, etc.
24.	20 July	Watched another 1/3rd of the previous day's lecture and learned about different types of data (raw, processed, and cooked) in the big data pipelines. Followed by their evolution — how flat file database stores data in plain text format, how in relational databases data are stored in tables, how the idea of data lake originated, and how the data pipelines property (with latencies, scalability, querying, and monitoring) came into the picture. Read through them diligently and called it a day.
25.	21 July	Began the day with watching the final portion of the same video, which explained about data warehouses in great detail. Then I read through a discussion about the differences between data warehouse and data lake, w.r.t data processing, storage, agility, security, and users. It was great, learning about the cost of storages, configuration modes, schema types, and data types. After that, I learned about different data pipeline solutions to automate the process end-to-end in an efficient, reliable, and secure manner. It focussed on batch, real-time, cloud native, and open source.
26.	22 July	22nd July was different. With all the stuff I'd learned thus far about data pipelines, now I'd relearn some of the concepts in order to implement them in the internet of things domain. The IoT Data Pipelines. Nonetheless, it was interesting. I read about data visualisation layer, data collection layer, data ingestion layer, data query layer engine, analytics engine, data processing layer, data storage layer, data security layer, and data monitoring layer. And their unique roles. I learned about several data sources as well.
27.	23 July	Started the day (and ended as well) with learning in brief about other technology stacks: Hadoop distributed file system, Spark streaming, Spark MLLib, Kafka, and Visualisation tools such as Tableau, Qlikview, D3.js, etc. And how they differ from MongoDB, in terms of availability, scalability, insulation, and consistency.
28.	24 July	Watched the final video of this course which explained in details how to build big data pipelines, using most of the technology stacks discussed earlier. It'd taken about 30 minutes. I also learned about different layers in it: speed, batch, and serving layers.
29.	25 July	Started just where I left off the previous day, and read about all the benefits of big data pipelines. I learned how it helps in better event framework designing, how data persistence is maintained, how it makes scalability very easy, and how it helps provide serialisation framework. I also learned about economic resources and how they may affect the performance of data pipelines, and

		how it leads to a violation of privacy on the cloud for critical data. And with this, the theoretical portion of the course ended.
30.	26 July	With the end of the major part of the course, I'd to submit the 3rd Assignment, based primarily on Module 3, before I could move forward. It was during the peak of India's 2nd wave of Covid-19, and the work I got was also based around it. First, I'd to show a practical example to list files, Insert data, retrieving data and shutting down HDFS. After that, building on the simple WordCount example done in class and Hadoop tutorial, my task was to perform simple processing on the provided COVID-19 dataset. The final task was to count the total number of reported cases for every country/location till April 8th, 2020. I'd to face a couple of problems while performing it, and the entire assignment from start to finish took away north of 6 hours of my day. However tedious, it was worth it.
31.	27 July	Being done with all the assignments, I took the day on the liberty of revising everything I'd studied on the course, starting from the introductory stuff of big data to the more complex Hadoop systems, other tech stacks, etc. Took me around 2 hours, and I called it a day.
32.	28 July	Took the Quiz. It'd a set of 25 questions. The result was out immediately, corrected the few mistakes I committed, and revised the entire course again, this time briefly. Then I took the Mock Assessment in lieu of checking my preparation for the Final Assessment Test.
33.	29 July	The last day. I gave the Final Exam of the course; it was scheduled for 60 minutes. And received my course completion badge on the Futureskills prime profile. The score card and course certificate were mailed to me (and being made available in my profile as well) in a few days' time.