HEMANT MUKHERJEE

Address *: L401,Kolte Patil Ragaa, Hennur, Bangalore*
e-mail *: hemantrec@gmail.com*
M *: +91-8295264809*

Result Oriented and Self Driven Data Engineering Professional with 6.4 years of Information Technology experience looking forward to senior data engineer roles in data-centric enterprises where I could contribute to value creation in terms of providing my expertise in solving real world data challenges in the business.

PROFILE SUMMARY

- A result oriented professional with 6.4 years of experience in Information Technology services industry focused on Big Data Engineering.
- Expertise in Big Data frameworks and tools – Apache Spark/Databricks on AWS, Apache Hadoop, Apache Hive, Apache Sqoop(Hortonworks and Cloudera Distributions),Snowflake
- Experience with machine learning frameworks such as MLlib,Scikit Learn,Numpy and other Machine Learning frameworks.
- Expert analytical and problem-solving skills; proficient in data-driven clustering, classification, ranking, and estimation techniques
- Experienced in designing , building, scheduling of scalable end to end data pipelines leveraging big data cloud tools.
- Experienced in application optimization and debugging techniques for improvement of overall time and quality of data ingestion.
- Worked efficiently in cross geographic and cross functional agile teams and demonstrated good communication and analytical skills enriched with exposure to team leading activities.

IT SKILLS

Programming Languages	Python , Scala and Core Java basics
Big Data Cloud Tools	Databricks, Apache Spark APIs(Core, SQL), Hadoop, HDFS, Yarn, Hive, MapReduce, Sqoop ,Sckit Learn,Numpy,Pandas,, Apache Airflow, AWS (EC2,EMR,S3,GLUE,SES)
Other Developer Tools	Putty, Eclipse IDE, IntelliJ IDE, PyCharm IDE, Maven,Jira, TD Platform, Rational Application Developer, ClearTeam Explorer, SecureCRT, SecureFx, WinSCP
Databases	MySQL, Oracle, Snowflake
Operating Systems	Windows, Unix, Linux, VM Platforms(Vmware, Citrix, Oracle VirtualBox)

*Upskilling	Apache Kafka, Apache Spark Streaming, AWS

EXPERIENCE

Duration	Organization	Designation	Location
08th March 2016 – 30th Aug2018	Harman Connected Systems	Associate Engineer Professional Services	Bangalore, Karnataka
19th Oct 2018 – 23rd Dec 2019	KPMG	Executive	Bangalore, Karnataka
30 th Dec2019- 24 July2020	Book My Show	Software Engineer	Bangalore, Karnataka
19th October 2020 – till now	EPAM Systems	Software Engineer	Bangalore, Karnataka

PROJECTS

#NBS MLDL: EPAM-Novartis Business Services Domain : Pharma/Life Sciences
Role : ML Engineer
Duration : March '22 – Till Date
Tools used : Python, Spark 3, Databricks on AWS, S3, Airflow, Snowflake, ML

Description:

Worked with Novartis Business Services as part of Use Case development team to develop and deliver a solution for a US region DW migration effort by leveraging Novartis' ML FormulaOne platform . The scope of the project had been to ingest the Commercial, Tech Ops and Finance & Supply Chain data from data sources into Data Lake and make predictions as per the business requirement.

Roles and responsibilities:

- Actively involved in analyzing and estimating on requirements and end to end data pipeline development and maintenance for 12 different data sources.(CSV files, Azure DB via SFTP, CRM)
- Responsible for creation of clusters in the data for model training and testing.
- Contributed to effective troubleshooting and bug fixing to maintain quality of the product.
- Responsible for model training and testing for the classification models using confusion matrix and AUC score..

- Taken active part in collaborating for POCs and code reviews with Novartis DE ML team as and when required.
- Helped in knowledge transition for newer project members and contributed to multiple project documentation targeted towards business and operational users of the software implementation.

#NBS F1DL: PDL EPAM-Novartis Business Services Domain : Pharma/Life Sciences
Role : Big Data Developer
Duration : July '21 – Mar'22
Tools used : Python, Spark 3, Databricks on AWS, S3, Airflow, Snowflake, AWS SES, Streamsets

Description:

Worked with Novartis Business Services as part of Use Case development team to develop and deliver a solution for a US region DW migration effort by leveraging Novartis' FormulaOne platform . The scope of the project had been to ingest the Commercial, Tech Ops and Finance & Supply Chain data from data sources into Data Lake and make it available for advanced analytics and reporting.

Roles and responsibilities:

- Actively involved in analyzing and estimating on requirements and end to end data pipeline development and maintenance for nine different data sources.(CSV files, Azure DB via SFTP, CRM)
- Responsible for creation and configuration of Pyspark jobs for the different data staging layers - raw , unification and publishing using Databricks notebooks on AWS and Snowflake. Also worked on scheduling the jobs through Airflow.
- Contributed to effective troubleshooting and bug fixing to maintain quality of the product. • Worked on creation of individual Python/Pyspark utilities to handle job concurrency issues, outbound file transfer back to source system and automated email reports for the business. • Responsible for Streamsets implementation in use case for CRM data source which enabled data ingestion for 60 + CRM objects into F1 platform as part of that effort.
- Taken active part in collaborating for POCs and code reviews with Novartis DE team as and when required.
- Helped in knowledge transition for newer project members and contributed to multiple project documentation targeted towards business and operational users of the software implementation.

#NBS F1DL: NBS-Digi(SFMC) EUData Warehouse
Domain : Pharma/Life Sciences **Role** : Big Data Developer **Duration** : Oct'20 – Apr'21
EPAM-Novartis Business Services

Tools used : Python, Spark 3, Databricks on AWS, S3, Airflow, AWS RDS, Streamsets

Description:

Worked with Novartis Business Services to develop and deliver a MVP for a cloud DW for the EU region leveraging the Novartis' new open, modular, secure, multi-tenant and industry compliant Cloud based scalable Data management & Analytics Platform named FormulaOne. The scope of the project had been to ingest the Commercial, Tech Ops and Finance & Supply Chain data from data sources into Data Lake and make it available for advanced analytics and reporting to support operations in EMEA.

Roles and responsibilities:

- Responsible for understanding the Novartis FormulaOne framework and requirement analysis for use cases to onboard 12 disparate data sources onto the platform.
- Involved in developing data pipelines for ingestion of data into raw layer of FormulaOne using PySpark through Databricks notebooks on AWS.
- Created scheduled jobs for multiple data sources in Airflow for raw as well as publish data layers
- Worked on publish layer data transformations to push data into Postgres DB hosted on AWS RDS instance which is directly consumed by business.
- Involved in collaborating with Novartis Data Engineering team on Data Quality Checks. • Contributed to effective troubleshooting and bug fixing to maintain quality of the product.

#BookMyshow : Entertainment
Role : Big Data Developer
Duration : Oct '19 – Jul'20
Tools used : Hortonworks, Hive , MapReduce ,Yarn ,HDFS, Apache Spark

Description:

I was part of a team which used SPARK with SCALA in Cloudera distribution to read the data stored in hive tables and analyze the data stored to find at which point the users are facing difficulties and then take business decisions.

Roles and responsibilities:

- Understood the Business Requirements as well the existing Data architecture supporting the application by coordinating with onshore counterparts and key project stakeholders on a daily basis.

- Understood and analyzed the nature of the metadata being worked upon.
- Worked on creation of Data pipelines and data ingestion for new enhancements and integration of same with frontend application using Hive as the data warehousing framework.
- Creation and maintenance of the HQL scripts, Java Mapper programs and various properties' files and working on performance tuning of applications and hive queries as and when required. • Worked on data processing using Apache Spark SQL over Hive Framework and storage of same in optimized file formats for consumption by downstream application.
- Also worked with RDD, DataFrame Spark APIs and have basic working knowledge of Spark Streaming. • Created clean, comprehensive design documents, data trend analysis reports and playbooks that were thoroughly appreciated by offshore and onshore teams and leads.
- Responsible for ensuring 100% code coverage in Unit Testing , and involvement in Test Data mock-up and Test Support.

#KPMG : Audit
Role : Data Engineer
Duration : Oct '19 – Dec '19
Tools used : Hortonworks, Sqoop, Hive, Yarn, HDFS

Description:

Worked with KPMG for data and analytics team Whose responsibility was to monitor and Business Intelligence solution for one of its major web channels through data analytics.kpmg.com is the main Interactive online portal targeted towards the e-business of the client. Primarily huge amounts of all registered customer data specific to their KPMGAccount are being processed and analyzed on our clusters on a weekly basis . Data cleansing and processing is undertaken by our team to aid in analysis of customer usage and behavior for implementation of targeted offers, targeted price and data plans, adverts, promotional and various business decisions.

Roles and responsibilities:

- Involved in daily calls involving various stakeholders (offshore and onshore) to understand the business requirements and agenda.
- Responsible in analyzing the customer requirements and application of appropriate solution. • Involved in creation and maintenance of Sqoop scripts and HQL scripts.
- Responsible in debugging of application and troubleshooting issues, and maintaining a detailed tracker about the issues.
- Worked with different file formats to optimize storage and transfer of data on and across cluster. • Involved in unit testing, test data management, test support and issue troubleshoot to ensure efficient QA practice release on release.
- Contributed to Application Knowledge documents, process documents and helped in knowledge transfer to new joiners .

EDUCATION

2015 : B.Tech in Computer Science Engineering from RGPV with 66.4 %

2009 : 12th in Science stream with 63%
2007 : ICSE with 72%

DECLARATION

I hereby declare that the information stated above is true to the best of my knowledge.

(Hemant Mukherjee)