Resume

Nitin Kumar

Sr. Data Engineer

Capable Data Engineer having 8.3 years of IT experience in Analysis, design, development, implementation, maintenance and support with experience in Big Data, Hadoop Development and Ecosystem Analytics, Development and Design of enterprise applications with expertise in Data Development in Cloudera and Hortonworks HDP platform & Hadoop ecosystem tools.

Contact

Address
New Delhi, Delhi, 110078

Phone
8178823113

E-mail
Dataengineer091990@gmail.com

LinkedIn
https://www.linkedin.com/in/nitin-kumar-a88ba248

Skills

- **Spark, PySpark**
- Big Data, Apache Hadoop
- Horton works, Cloudera distribution
- **Hive**, Pig, Sqoop, Oozie, Kafka, KSQL
- Nosql – **Casssandra, Hbase**

Work History

2020-09 – Current

Sr. Data Engineer/Assistant Vice President
Citi Corp, Pune, Maharashtra
GFTS Consumer Application

- Working on creating generic framework for data ingestion from various sources like HDFS (Csv, Xml, Json), Hbase, JDBC (Netezza and Oracle), Kafka Topic etc and after applying some transformations dumping data on various sink system as per the application requirement.
- Business requirement gathering from various application teams to incorporate in the framework being developed.
- Writing junit test cases for the various sources and sinks classes written.
- Writing similar functionality in python as well for the ingestion framework so it could be more extensible if machine learning is also needed to integrate for any application team.
- Creating tableau subscription board to demonstrate the volume pattern change.

- ETL Development, SQL, MySQL, Oracle
- **AWS Services, Dynamo DB, S3, Redshift, EMR etc**
- **Machine Learning – Linear Regression, Logistics**
- Unix Scripting
- **Tableau**
- Airflow, Autosys, Event Engine
- Java, **Python**, Scala

Domains

- Travel
- Retail
- Banking & Financial Services

2019-01 – 2020-09

Sr. Data Engineer/Senior Consultant

Cognizant Technologies, Gurgaon, Haryana

Client - American Express – Global Commercial Services (Jan 2020 - Ongoing)

- Design Architecture of data pipeline/ingestion as well as optimization of ETL workflows and developed curriculum data pipelines from Syllabus/Curriculum Web Services to Cassandra and Hive tables.
- Performed data analysis, feature selection, feature extraction using Apache Spark libraries in Scala, Python and Developed data pipeline for real time use cases using Kafka, Flume and Spark Streaming.
- Working on Cloud computing using Amazon Web Services various BI Technologies and exploring NoSQL options for current back using AWS Dynamo DB (SQL API)
- Enable and configure Hadoop services such as HDFS, YARN, Hive, Hbase, Kafka, Sqoop, Zeppeline Notebook and Spark and involved in analyzing log data to predict the errors by using Apache Spark.
- Expertise in Data Development in Cloudera and Hortonworks HDP platform &Hadoop ecosystem tools like Hadoop, HDFS, Spark, Hive, HBase, SQOOP, flume, Pig, Oozie, Hue, Tez, Kafka.
- Extracting real time data using Kafka and spark streaming by Creating DStreams and converting them into RDD, processing it and stored it into Cassandra.
- Created Hive tables, loaded data and wrote Hive queries that helped market analysts spot emerging trends by comparing fresh data with EDW reference tables and historical metrics.
- Used Hive QL to analyze the partitioned and bucketed data and compute various metrics for reporting and performed data transformations by writing MapReduce and Pig jobs as per business requirements

Client - American Express – Credit and Fraud Risks Capabilities (Jan 2019 - Jan 2020)

- Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark SQL, Data Frame, and Spark Yarn.
- Worked on cloud computing infrastructure (e.g. Amazon Web Services EC2) and considerations for scalable, distributed systems and involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS and converted all Hadoop jobs to run in EMR by configuring the cluster according to the data size.
- Used AWS Data Pipeline to schedule an Amazon EMR cluster to clean and process web server logs stored in Amazon S3 bucket.
- Predictive modeling experience in using logistic regression to predict binary outcomes for credit risk analysis with the use of SAS.
- Worked on data pipeline creation to convert incoming data to a common format, prepare data for analysis and visualization, Migrate between databases, share data processing logic across web apps, batch jobs, and APIs, Consume large XML, CSV, and fixed-width files and created data pipelines in kafka to Replace batch jobs with real-time data.
- Developed data pipeline using Spark, Hive and HBase to ingest customer behavioral data and financial histories into Hadoop cluster for analysis.
- Worked on CICD Automation using tools like Jenkins, Swagger Git, Code deploy.
- Data Management, Data Access, Data Govern ance and Integration, Security, and Operatons performed by using Hortonworks Data Platform (HDP).
- Worked with importing metadata into Hive using Python and migrated existing tables and applications to work on AWS cloud (S3).

2016-03 - 2019-01

Sr. Big Data Developer

Tata Consultancy Services, Noida, Uttar Pradesh

Client - Apple GBI iTunes Financial (Jan 2018 - Jan 2019)

- Developed UNIX scripts in creating Batch load for bringing huge amount of data from Relational databases to BIGDATA platform.
- Delivery experience on major Hadoop ecosystem Components such as Pig, Hive, Spark Kafka, Elastic Search & HBase and monitoring with Cloudera Manager.
- Implemented Kafka consumers to move data from Kafka partitions into Cassandra for near real-time analysis and worked extensively on Hive to create, alter and drop tables and involved in writing hive queries.
- Involved in requirement and design phase to implement Streaming Architecture to use real time streaming using Spark and Kafka.
- Implemented the Business Rules in Spark/ SCALA to get the business logic in place to run the Rating Engine.
- Extensively used ETL methodology for supporting Data Extraction, transformations and loading processing, using Hadoop.
- Used both Hive context as well as SQL context of Spark to do the initial testing of the Spark job and used WINSCP and FTP to view the data storage structure in the server and to upload JARs which were used to do the Spark Submit.

Client - Apple iTunes Enterprise Semantic Layer (ESL) Billing (Mar 2016-Jan 2018)

- Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
- Uploaded and processed terabytes of data from various structured and unstructured sources into HDFS using Sqoop.

- Created Reports with different Selection Criteria from Hive Tables on the data residing in Data Lake.
- Developed workflow in Oozie to automate the tasks of loading the data into HDFS and pre-processing with Pig and parsed high-level design spec to simple ETL coding and mapping standards.
- Designed and developed exception handling, data standardization procedures and data quality assurance controls.
- Implemented the Machine learning algorithms using Spark with Python and worked on Spark Storm and python.
- Involved in configuring batch job to perform ingestion of the source files in to the Data Lake and developed Pig queries to load data to HBase
- Involved with the team of fetching live stream data from DB2 to Hbase table using Spark Streaming and Apache Kafka.
- Extensively used Stash Git-Bucket for Code Control and Worked on AWS Components such as Airflow, Elastic Map Reduce (EMR) and Snow-Flake.

2012-12 - 2016-03

Software Engineer

InterGlobe Technologies Pvt Ltd, Gurgaon, Haryana

Client - Travelport GWS Select (Feb 2014 – March 2016)

- Involved in gathering and analyzing system requirements and played key role in the high-level design for the implementation of this application.
- Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
- Developed the application using Eclipse IDE and worked under Agile Environment and

worked with Web admin and the admin team to configure the application on development, training, test and stress environments (Web logic server).
- Created the WSDL and restful services for publishing the WSDL and creating PDF files for storing the data required for module.
- Used spring framework configuration files to manage objects and to achieve dependency injection.
- Implemented CI, CD using Jenkins for continuous development and delivery.
- Extensively worked on TOAD for interacting with data base, developing the stored procedures and promoting SQL changes to QA and Production Environments.

Client - Travel Port E-Ticket Management (Dec 2012 - Feb 2014)

- Imported data from RDBMS systems like MySQL into HDFS using Sqoop and developed Sqoop jobs to perform incremental imports into Hive tables.
- Wrote POWERSHELL scripts to copy or move data from local file system to HDFS Blob storage.
- Involved in creating Data Lake by extracting customer Big Data from various data sources into Hadoop HDFS. This included data from Excel, Flat Files, Oracle, SQL Server, Cassandra, HBase
- Coordinated with the external teams to assure the data quality of master data and conduct UAT/integration testing
- Delivery experience on major Hadoop ecosystem Components such as Pig, Hive, Elastic Search & HBase and monitoring with Cloudera Manager.

Education

2008-05 - 2012-08
78%

B. Tech (IT): Information Technology

Delhi Institute of Technology And Management, MDU - Sonepat, Haryana

2007-04 - 2008-03
81%

Intermediate: Science/Computers

MCL Sarawati Bal Mandir - Delhi
CBSE Board

2005-04 - 2006-03
83%

High School: Science

Saraswati Bal Mandir - Delhi
CBSE board.

Certifications

- OCPJP (Oracle Certified Professional for Java Programmer) Ver-6 Certified.
- Udemy Certified AWS Developer Associate.
- Udemy Certified Apache Cassandra Developer.