HETAL BHADAJA

Senior Software Engineer

New Jersey 07054 • 973-864-0314 • hetal0606.ds@gmail

LinkedIn: www.linkedin.com/in/hetal-bhadaja-400bb7154

Summary

· AWS (S3, Redshift, EC2)	· Data Bricks
· Spark (scala, sql)	· Qubole
· Visual studio	· Alteryx
· Sql Server, Oracle9i/10g, DB2, Teradata	· GIT
· Tableau	· Agile/Scrum Methodology

- Worked on Spark Architecture including spark Core, Spark SQL, Data Frame, Spark Streaming.
- Hands on development on AWS platform with EC2, S3 & EMR
- Experience working with AWS security and managing IAM roles
- Developed an end-to-end data analytics framework utilizing Amazon Redshift and various sources.
- Designed and modeled Hive database using Partitioned and Bucketing tables with storing data in various file systems like Parquet, RC, ORC and Text File
- Scheduled importing and exporting batch data using Sqoop and real-time data using Flume from Relational Database Systems to HDFS and vice-versa
- Experience in monitoring infrastructure for the Spark cluster using Ganglia
- Worked using Integrated Development environments like Eclipse, Intellij IDEA and Version Control tools like SVN, GIT and Build tools like Maven and SBT
- Worked and migrated RDBMS databases into different NoSQL database
- In depth exposure to HDFS concepts, blocks, master nodes, slave nodes, HDFS high-availability including both manual and automatic failover, command line interface, file reading and writing, copying a file from one cluster to another
- Designed and implemented a complete end-to-end Big Data solution using Hadoop Ecosystem tools, including HDFS, MapReduce, Hive, Pig, Sqoop and Spark
- Performed Infrastructure capacity planning considering various workload patterns and actively participated in hardware design for Master and Worker Nodes
- Experience in installing and configuring Hadoop clusters using different distributions of Apache Hadoop like Cloudera and EMR
- Analyzed the clients existing Cluster infrastructure and understand the performance bottlenecks and provide the performance tuning accordingly
- Experience in PL/SQL programming including SQL queries using stored procedures and triggers in Oracle, SQL Server using TOAD and Query Manager
- Experience in integration of various data sources like Oracle9i/10g, DB2, Teradata, SQL Server, MS Access, XML and Flat files
- Extensively worked on leading ETL design, development, testing and implementation phases. Created logical and physical mapping documents and low level/high level ETL design documents using InfoSphere Information server business suites.
- Worked within DevOps practices throughout the application lifecycle which help accelerate, automate, and improve a specific phase.

EXPERIENCE

PUBLICIS MEDIA, New York, NY **07/2018 – 12/2023**

Senior Software Engineer

Project Description:

Publicis Media is one of Publicis Groupe's four solution hubs, aligning all of Publicis Groupe's media agencies and operations. Publicis Groupe is the world's third largest communications group. The data, technology, and innovation global practice was created to deliver best-in-class programmatic solutions as well as to consolidate Publicis Media's data and technology to transform the business from a service business to a platform business. The project was to design, develop, and implement and on-going support of data processing systems and larger data platforms.

Responsibilities:

- Designed, built and maintained efficient, reusable, and reliable ETL processes by the data processing framework within latest Cloud Technologies to provide analytics to the reporting team for business intelligence dashboards.
- Worked with large digital media files, including impression, click and activity files, as part of the analytics process for media reporting.
- Clearly understand spark architecture and find the best solution to work within instead of various systems.
- Create scala written code to access millions of impression files and transform it as per analytics team requirement.
- Implement solution within spark scala partition which saves query runtime and resources. which helps to meet SLA
- Experience in spark cluster tanning which helps company to save in finance.
- Performed capacity planning for the Development, Test and Production environments to inform the budget planning process that enabled the right amount of capacity on the system.
- Extensively used Qubole Interface initially and then migrate work at Databricks.
- Setup Databricks console from scratch as administrator and create users, groups and assign them roles. Create tables and schemas within Unity Catalog concept and did POC for Delta live table i.e bronze, silver and gold tables
- Architected and Designed Spark cluster of up to 30 nodes consisting of Master nodes and Worker nodes for efficiency
- Configured Sqoop Client which allow us to import data from multiple databases into HDFS to make deriving analytics more efficient
- Design the Unity Catalog on Databricks as part of a larger team
- Wrote spark jobs in Scale to read large data sets in text/parquet file formats from data lake by converting into RDDs, Data frames and Datasets with custom logic and various Spark APIs to enrich the data as per business needs.
- Created and Managed storage format strategies and Data Partitioning and Bucketing strategies for Hive tables to create easier access for other teams.
- Designed and Developed Data Pipelines (ETL processes) in Apache Spark using Spark Context, Spark SQL and Spark Streaming applications using various techniques like Broadcast Variable, RDD Lineage, Caching and Distributed persistence, Accumulators, Kryo Serialization, Repartition/Coalesce
- Extensively built an email notification system with AWS SES and other tools as per the requirement for QA and other alerts

- Developed Alteryx job to integrate different sources like databases, S3, excel with ETL and load to one place for Analytics reporting.
- Did setup Databricks console and migrate all work from Qubole
- Identified bottlenecks and bugs, and devise solutions to mitigate and address these issues

Environment:
*Amazon Web Services like S3,EC2, Qubole Interface, Presto, Sql Server DB, Spark Scala , Redshift, Databricks Interface, T*ableau, Alteryx and Visual Studio, JIRA Reporting tool

SOFTWARE CONSULTING GROUP, **PA** **11/2012 – 06/2018**

JETBLUE AIRWAYS, New York City, NY (03/2018 – 06/2018)

Spark Data Engineer

Project Description:
JetBlue Airways Corporation, stylized as JetBlue, is the sixth-largest airline in the United States. It was established in 2000 and is headquartered in Long Island City, New York, NY. The project was to build and migrate trillions of data from traditional databases to Cloud through Spark.

Responsibilities:
- Design and Develop Apache Spark (Python and Scala) to analyse trillions of data points
- Design and Develop Data Pipelines to ingest billions of data points daily with Kafka and Spark Streaming
- Build very large database solutions to support new app feature
- Implement data solution on Microsoft Cloud(Azure) using various Azure services
- Monitor and Orchestrate Hopper service using Kubernets and Mesos
- Design and Develop Restful scala-based services for distributed systems
- Seamlessly Integrate data analysis solution with large distributed no-sql database on cloud
- Working in Development environments like Intellij IDEA and Version Control tools GIT HUB and Build tools Maven and SBT
- Working thorough in Scala to read and process large data sets in json file formats by converting into RDDs, Data frames and Datasets with custom logic to enrich the data as per business needs
- Analyzing the clients existing Cloud cluster, understand the performance bottlenecks and provide the performance tuning accordingly

Environment:
Microsoft Azure with Cosmos DB, Blob, HDInsight Spark Cluster, and Databricks.

New York City DHS, New York City, NY (08/2015 – 02/2017)

Senior Developer

Project Description:

Since 1993, DHS has been one of the largest organizations of its kind committed to preventing and addressing homelessness in New York City. The mission of Homeless Services is to overcome homelessness inNew York City. The project was to build Enterprise Data Warehouse for integrating data from all heterogeneous sources and build DataMart for Shelter and Client Information and generate required monthly files. Run report as per request.

Responsibilities:

- Architect, Design and Develop all ETL components including data acquisition, cleansing, standardization, profiling, validation, staging and database persistence using IBM InfoSphere suite, especially InfoSphere Datastage and Information Analyzer.
- Extensively worked in converting Business Requirements into Functional and Non-Functional (Technical) Specifications.
- Expertise in using Object Oriented, Entity Relationship, Dimensional Modeling techniques.
- Provided Technical expertise in solving complex data Integration issues.
- Architected and Designed Hadoop cluster for POC with up to 8 nodes consisting of Edge nodes, Management nodes and Data nodes.
- Used Cloudera Manager to manage cluster nodes, services, administering cluster and assigning users, groups and roles for authorization.
- Set up and configured SFTP servers to import data from third party vendors.
- Designed and Developed Data ingestion process from various sources using Apache Sqoop, Flume, Hadoop shell commands and Kafka.
- Designed Data warehouse database on Hive.

Environment:

IBM / Ascential DataStage 8.7,8.5, Oracle 11g, PostgreSql, MS Word ,MS Access, Unix Windows 2007,Cognos 10.2, Visio 2013, JIRA Reporting tool, Cloudera Enterprise CDH 5.4, Hadoop 2.6, Hive 1.1, Sqoop 2.0, Spark 1.3,Aginity Workbench

NEW JERSEY DEPARTMENT OF TRANSPORTATION, Trenton, NJ (03/2014 – 07/2015)

Software Developer

Project Description:

NJDOT was established in 1966 as the first state transportation agency in the united states. It has been responsible for maintaining and operating the state's highway and public road system, planning and developing transportation policy, and assisting with rail, freight, and intermodal transportation issues. The project was to build an enterprise data warehouse for integrating data from all heterogeneous sources and build data mart for transit information.

Responsibilities:

- Worked closely with the ETL Lead, Technical Lead, Data Modeler, Business Analysts to understand business requirements, providing expert knowledge and solutions on Data Warehousing, ensuring delivery of business needs in a timely cost-effective manner
- Based on Technical Specification Document designed and created Datastage jobs to extract, transform and load data from source into Data warehouse and then into Datamarts with different partitioning methods like Hash by field, Round Robin, Entire, Modulus, and Range for bulk data loading

- Deployed Implemented Slowly Changing Dimension Type- 2 in Datastage
- Extensively used DataStage Designer stages such as Aggregator, Transformer, Join, Dataset, Lookup, Funnel, Peek, Pivot
- Lead the other team members and guided the testing team members in executing the jobs, which helped them to develop their knowledge on the tool
- Responsible for data analysis, requirements gathering, source-to-target mapping, process flow diagrams, and documentation
- Created UNIX shell scripts for File Transfer and File validation during parallel job execution
- Tuned SQL queries for better performance for processing business logic in the database

Environment:
IBM/Ascential DataStage 8.7, 8.5, Oracle 11g, SQL Server 2008, MS Word and Access, Unix Windows 2007, and Erwin 4.1.

KAISER PERMANENTE, Pleasanton, CA (11/2012 – 01/2014)

DataStage and Technical Lead

Project Description:
Kaiser Permanente is America's leading nonprofit integrated health plan. Kaiser Permanente is an integrated managed care consortium based in Oakland, CA, United States. The project is specially for performance tuning of most imported Datastage jobs which are missing SLA's

Responsibilities:
- Analyzed the Data Acquisition, Data Integration, Data Transformation, and Data Delivery processes for Performance Assessment.
- Created Scope document for Performance Assessment.
- Identified opportunities of improvement in Overall design, Installation, Configuration, Hardware, Network, Database, Scheduling, Strategy and Code.
- Created Hourly, Daily and Weekly runtime metrics to identify long running processes.
- Modified existing ETL and BI architecture to achieve maximum parallel processing and reducing overhead on Database servers.
- Changed scheduling of ETL and BI processes to run maximum processes in parallel.
- Tuned Database queries using various Oracle tuning techniques.
- Partitioned and Sub-partitioned tables choosing appropriate partition keys in Oracle.
- Analyzed Job Monitor, Score Dump, Resource Estimation and Performance Analysis to identify bottlenecks in Datastage jobs.
- Redesigned Datastage jobs to achieve better performance.
- Choose proper data partitioning and sorting methods to get optimum performance in Data stage.
- Lead the Design, Development, Testing and Implementation of performance improvement recommendations.
- Successfully improved Nightly process from Twelve hours to Four Hours.
- Provided strategic guidance and technical oversight during an engagement.
- Established Data Warehouse, Data Modeling, ETL and BI Standards and Best Practices.

Environment:

IBM/Ascential DataStage 9.1/8.5, Database Oracle 11g, MS Access, SQL, Unix, Linux, and Windows 2007.

PSP CONSTRUCTION INC., Hackensack, NJ **09/2011 – 10/2012**

Software Engineer

- Established and maintained tendering processes and setup cost monitoring and reporting systems and procedures
- Create Project Plan with related Software like Primavera and Microsoft Project
- Value engineering studies, constructability, design and specifications reviews
- Evaluated and reported on the progress control of projects
- Coordinate with Subcontractors, architects and agencies.

EDUCATION

Bachelor of Engineering (B.E.), **SAURASTRA UNIVERSITY**, India
Master of Engineering (M.E.), **CITY COLLEGE OF NEW YORK**, NYC

CERTIFICATION

ACADEMY ACCREDITATION - DATABRICKS LAKEHOUSE FUNDAMENTALS

973-864-0314 hetal0606.ds@gmail