**Name: Datta B**

**dattab2208@gmail.com**

**908-617-1396**

**Data Engineer**

**SUMMARY:**

* Having around 8 years of total IT experience with over 5 years’ experience in Big Data Hadoop experience in Development and Design of Java based enterprise applications.
* Extensive working experience on Hadoop eco-system components like **Hadoop,** **HDFS, Map Reduce, Hive, Sqoop, Flume, Spark, Kafka, Oozie and Zookeeper**.
* Implemented performance tuning techniques for Spark-SQL queries.
* Strong knowledge on Hadoop HDFS architecture, **Map-Reduce (MRv1) and YARN (MRv2)** framework.
* Strong hands on Experience in publishing the messages to various **Kafka** topics using **Apache NIFI** and consuming the message to HBase using **Spark and Python**.
* Experience in Developing Spark applications using **Spark - SQL in Data bricks** for data extraction, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Good understanding of Spark Architecture including **Spark Core, Spark SQL, Data Frames, Spark Streaming,** Driver Node, Worker Node, Stages, Executors and Tasks.
* Experience with MS SQL Server Integration Services (SSIS), T-SQL skills, stored procedures, triggers.
* Design and develop Spark applications using Pyspark and Spark-SQL for data extraction, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Design and implement database solutions in Azure SQL Data Warehouse, Azure SQL
* **Azure Data Factory** (ADF), Integration Run Time (IR), File System Data Ingestion, Relational Data Ingestion.
* Worked on creating Spark jobs that process the true source files and successful in performing various transformations on the source data using **Spark Data frame, Spark SQL API's**.
* Developed Sqoop scripts to migrate data from **Teradata, Oracle** to **Big data** Environment.
* Experience in importing and exporting the data using Sqoop from HDFS to Relational Database systems and vice-versa.
* Hands on experience in installation, configuration, supporting and managing Hadoop Clusters using Apache, **Cloudera (CDH3, CDH4), Yarn distributions (CDH 5.X).**
* Implemented real time data streaming pipeline using **AWS Kinesis, Lambda,** and **Dynamo DB** and deployed **AWS Lambda** code from Amazon **S3** buckets.
* Work on large scale data transfer across different Hadoop clusters, implement new technology stacks on Hadoop clusters using Apache Spark.
* Added support for AWSS3 and RDS to host static/media files and the database into Amazon Cloud.
* Experience in project deployment using Heroku/Jenkins and using web services like Amazon Web Services (AWS) EC2, AWS S3, Auto scaling, Cloud Watch and SNS.
* Performed Data scrubbing and processing with Oozie and for workflow automation and coordination.
* Hands on experience in analyzing log files for Hadoop and eco-system services and finding root cause.
* Hands on experience on handling different file formats like AVRO, PARQUET, Sequential files, MAP Files, CSV, xml, log ORC and RC.
* Experience with NoSQL Database HBase, Cassandra, MongoDB.
* Experience with AIX/Linux RHEL, UNIX Shell Scripting and SQL Server 2008.
* Worked on data search tool Elastic Search and data collection tool Logstash.
* Strong knowledge in Hadoop cluster installation, capacity planning and performance tuning, benchmarking, disaster recovery plan and application deployment in production cluster.
* Experience in developing stored procedures, triggers using SQL, PL/SQL in relational databases such as MS SQL Server 2005/2008.
* Exposed into methodologies Scrum, Agile and Waterfall.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Programming Languages | Java, Python, SQL, and C/C++ |
| Big Data Ecosystem | Hadoop, Map Reduce, Kafka, Spark, Pig, Hive, YARN, Flume, Sqoop, Oozie, Zookeeper, Talend. |
| Hadoop Distributions | Cloudera Enterprise, Horton Works, EMC Pivotal. |
| Databases | Oracle, SQL Server, PostgreSQL. |
| Web Technologies | HTML, XML, JQuery, Ajax, CSS, JavaScript, JSON. |
| Streaming Tools | Kafka |
| Testing | Hadoop Testing, Hive Testing, MRUnit. |
| Operating Systems | Linux Red Hat/Ubuntu/CentOS, Windows 10/8.1/7/XP. |
| Cloud | AWS EMR, Glue, RDS, Cloud Watch, S3, Redshift Cluster, Kinesis, Dynamo DB. |
| Technologies and Tools | Servlets, JSP, Spring (Boot, MVC, Batch, Security), Web Services, Hibernate, Maven, GitHub, Bamboo. |

**Professional Experience: .**

**Client: NewYork Life Insurance, location: Lebanon,NJ Dec 2022 to till date**

**Role: AWS Data Engineer**

**Roles & Responsibilities:**

* Developed PySpark Applications by using python and Implemented Apache PySpark data processing project to handle data from various RDBMS and Streaming sources.
* Handled importing of data from various data sources, performed data control checks using PySpark and loaded data into HDFS.
* Involved in converting Hive/SQL queries into PySpark transformations using Spark RDD, python.
* Used PySpark SQL to Load JSON data and create Schema RDD and loaded it into Hive Tables and handled structured data using Spark SQL.
* Developed PySpark Programs using python and performed transformations and actions on RDD's.
* Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDD's.
* Used PySpark and Spark SQL to read the parquet data and create the tables in hive using the python API.
* Implemented PySpark using python and utilizing Data frames and PySpark SQL API for faster processing of data.
* Developed python scripts, UDFs using both Data frames/SQL/Data sets and RDD/Map Reduce in Spark 1.6 for data Aggregation, queries and writing data back into OLTP system through Sqoop.
* Experienced in handling large datasets using Partitions, PySpark in Memory capabilities, Broadcasts in PySpark, effective & efficient Joins, Transformations and other during ingestion process itself.
* Processing the schema oriented and non-schema-oriented data using python and Spark.
* Involved in writing live Real-time Processing and core jobs using Spark Streaming with Kafka as a data pipe-line system.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in HDFS.
* Worked on streaming pipeline that uses PySpark to read data from Kafka, transform it and write it to HDFS.
* Designed Data Marts by following Star Schema and Snowflake Schema Methodology, using industry leading Data Modeling tools.
* Worked on Snowflake database on queries and writing Stored Procedures for normalization.
* Worked with Snowflake’s stored procedures, used procedures with corresponding DDL statements, used JavaScript API to easily wrap and execute numerous SQL queries.

**Client: Impetus Technology, location: Los Gatos,CA Aug 2022 to Dec 2022**

**Role: Azure Data Engineer**

**Roles & Responsibilities:**

* Design & implement Migration strategies from SAS-Environment to Azure Cloud.
* Developing spark applications using spark-SQL in databricks.
* Develop spark applications using pyspark and spark SQL for data extraction, Transformation and Aggregation from multiple file formats for analzing and transforming the data uncover insight into the customer usuage patterns.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/to extract, transform and load data from different Sources like Azure SQL, Blob Storage.
* Experienced in Performance Tuning of spark aplications, correct level of parallelism and memory tuning.
* Developed JSON Scripts for deploying the Pipeline in Azure Data factory that process the data using the SQL activity.
* Interacts with Business analyst, users and SME’s on elobrating the requirement.
* Stage the API or kafka data(in JSON file format) into snowflake DB by FLATTENing the same for different funcational services.

**Client: Anthem, location: Virginia Beach, Virginia Nov 2019 to till July 2022**

**Role: Azure Data Engineer**

**Roles & Responsibilities:**

* Build Data Pipleline Architecture on Azure Cloud platform using NiFi, Azure DataLake Storage Service, Azure HD Insight , Airflow and Data Engineer tool.
* Designed and developed scalable and cost-effective architecture in Azure Big Data services for data life cycle of collection, ingestion, storage, processing, and visualization.
* Design and implement database solutions in Azure SQL Data Warehouse, Azure SQL.
* Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB).
* Design & implement migration strategies for traditional systems on Azure (Lift and shift/Azure Migrate, other third-party tools.
* Engage with business users to gather requirements, design visualizations and provide training to use self-service BI tools.
* Used various sources to pull data into Power BI such as SQL Server, Excel, Oracle, SQL Azure etc.
* Propose architectures considering cost/spend in Azure and develop recommendations to right-size data infrastructure.
* Develop conceptual solutions & create proof-of-concepts to demonstrate viability of solutions.
* Migrate ETL mappings and procedures across environments- development UAT/SIT and eventually to production and building migration document.
* Identify and implement best practices, tools and standards.
* Design Setup maintain Administrator the Azure SQL Database, Azure Analysis Service, Azure SQL Data warehouse, Azure Data Factory, Azure SQL Data warehouse.
* Build Complex distributed systems involving huge amount data handling, collecting metrics building data pipeline, and Analytics.
* Create SQL,PL/SQL scripts for sourcing data,including creating tables Materialized views,stored procedures and loading data into the tables.
* Used Power Bi,Power Pivot to develop data analysis prototype,and used Power View and Power Map to visualize reports.
* Published Power BI reports in the required originations and made Power BI Dashboards in Web clients and mobile apps.
* Implementing the Data Quality and content Validation by using tools like Spark,Scala,Hive,Nifi.
* Involved in creating End-to-End data pipeline within distributed environment using the Big data tools, Sparkframework and Power BI for data visualization.
* Redesigned the views in snowflake to increase the performance.
* Unit tested the data between Redshift and Snowflake.
* Developed data warehouse model in snowflake for over 100 data sets using wherespace.

**Client: Vanguard, location: Valley Forge, PA Dec 2016 TO Nov 2019**

**Role: Big Data Developer**

**Roles & Responsibilities:**

* Implementing the Proof of Concept (POC) for ETL Abinitio graph concepts which need to be migrated into Spark using scala and python (Pyspark).
* Develop Data pipelines using Sqoop, Spark, and Map reduce and Hive to Ingest, transform and analyze customer behavior data.
* Developed a data pipeline using Kafka, Spark Streaming and Hive to ingest the data from data lakes to Hadoop distributed file system.
* Implemented Spark using python and Spark SQL for faster processing of data and algorithms for real time analysis in Spark.
* Used Spark for interactive queries, processing of streaming data and integration with popular NOSQL database for huge volume of data.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Python.
* Handled importing data from different data sources into HDFS using Sqoop and also performing transformations using Hive, Map Reduce and then loading data into HDFS.
* Extracting the Files from the RDBMS (DB2) by using Sqoop into Hadoop file system (HDFS) to process the workflow.
* Implementing the Partitioning and bucketing for faster query processing in Hive Query Language (HQL).
* Involving in Converting the HIVE/SQL queries into Spark transformation using Spark Data frames, Datasets and User defined functions (UDF's).
* Design Hive queries and Pig Script to perform Data Analysis, Data transfer and Table design.
* Evaluating the Data between the ETL and Hadoop to ensure Data quality.
* Responsible in creating mappings and workflows to extract and load data from relational databases, flat file sources and legacy systems.
* Testing on Apache Tez Framework and Hadoop Map Reduce frameworks for building high performance batch and interactive Data Processing Applications.
* Reconciling the data on daily basis in between the ETL and Hive tables by using a compare tool which is implemented in spark framework using Pyspark.
* Fine tune Hadoop applications for high performance and throughput, troubleshoot and debug any Hadoop ecosystem run time issues.
* Performing Data validation operation between ETL and Apache Hive tables.
* Developing the Linux shell scripting for Deploying and running the migrated Hadoop Application in Production Servers.
* Developing Workflows for scheduling and orchestrating the Hadoop Process.

**Client: Accenture, Location: Hyderabad, India March 2014 to DEC 2015**

**SQL Developer**

**Roles & Responsibilities:**

* Research and recommend suitable technology stack for Hadoop migration considering current enterprise architecture.
* Extract Transform Load (ETL) development using SQL SERVER 2008,SQL 2012 Integration services(SSIS)
* Extensively used Spark stack to develop preprocessing job which includes RDD, Datasets and Data frames Api's to transform the data for upstream consumption.
* Developed Real-time data processing applications by using Scala and Python and implemented Apache Spark Streaming from various streaming sources like Kafka, Flume and JMS.
* Replaced the existing Map Reduce programs into Spark application using Scala.
* Built on premise data pipelines using Kafka and Spark streaming using the feed from API streaming Gateway REST service.
* Developed the Hive UDF's to handle data quality and create filtered datasets for further processing
* Experienced in writing Sqoop scripts to import data into Hive/HDFS from RDBMS.
* Good knowledge on Kafka streams API for data transformation.
* Developed oozie workflow for scheduling & orchestrating the ETL process.
* Used Talend tool to create workflows for processing data from multiple source systems.
* Created sample flows in Talend, Stream sets with custom coded jars and analyzed the performance of Stream sets and Kafka steams.
* Tested Apache TEZ, an extensible framework for building high performance batch and interactive data processing applications, on Pig and Hive jobs
* Optimized Hive QL/ pig scripts by using execution engine like Tez, Spark.
* Developed Hive Queries to analyze the data in HDFS to identify issues and behavioral patterns.
* Involved in writing optimized Pig Script along with developing and testing Pig Latin Scripts.
* Deployed applications using Jenkins framework integrating Git- version control with it.
* Participated in production support on a regular basis to support the Analytics platform
* Used Rally for task/bug tracking.
* Used GIT for version control.