Jessica Claire

100 Montgomery St. 10th Floor (555) 432-1000 resumesample@example.com

SUMMARY

- Dynamic and motivated IT professional with around 7 years of experience as a Big Data Engineer with expertise in designing data intensive applications using Hadoop Ecosystem , Big Data Analytical , Cloud Data engineering , Data Warehouse / Data Mart, Data Visualization , Reporting , and Data Quality solutions .
- In - depth knowledge of Hadoop architecture and its components like YARN , HDFS, Name Node, Data Node, Job Tracker, Application Master, Resource Manager , Task Tracker and Map Reduce programming paradigm.
- Extensive experience in Hadoop led development of enterprise level solutions utilizing Hadoop components such as Apache Spark, MapReduce, HDFS, Sqoop, PIG, Hive, HBase, Oozie, Flume, NiFi, Kafka, Zookeeper, and YARN.
- Profound experience in performing Data Ingestion, Data Processing (Transformations, enrichment, and aggregations).
- Strong Knowledge on Architecture of Distributed systems and Parallel processing, In-depth understanding of MapReduce programming paradigm and Spark execution framework.
- Experienced with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context , Spark-SQL , Dataframe API , Spark Streaming, MLlib , Pair RDD 's and worked explicitly on PySpark and Scala .
- Handled ingestion of data from different data sources into HDFS using Sqoop, Flume and perform transformations using Hive, Map Reduce and then loading data into HDFS .
- Managed Sqoop jobs with incremental load to populate HIVE external tables. Experience in importing streaming data into HDFS using Flume sources, and Flume sinks and transforming the data using Flume interceptors.
- Experience in Oozie and workflow scheduler to manage Hadoop jobs by Direct Acyclic Graph (DAG) of actions with control flows.
- Implemented the security requirements for Hadoop and integrating with Kerberos authentication infrastructure- KDC server setup, creating realm /domain, managing.
- Experience of Partitions, bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance .
- Experience with different file formats like Avro , parquet , ORC , Json and XML .
- Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow and Oozie.
- Experienced with using most common Operators in Airflow - Python Operator, Bash Operator, Google Cloud Storage Download Operator, Google Cloud Storage Object Sensor.
- Hands-on experience in handling database issues and connections with SQL and NoSQL databases such as MongoDB , HBase , Cassandra , SQL server , and PostgreSQL .
- Created Java apps to handle data in MongoDB and HBase. Used Phoenix to create SQL layer on HBase.
- Experience in designing and creating RDBMS Tables, Views, User Created Data Types, Indexes, Stored Procedures, Cursors, Triggers and Transactions .
- Expert in designing ETL data flows using creating mappings/workflows to extract data from SQL Server and Data Migration and Transformation from Oracle/Access/Excel Sheets using SQL Server SSIS .
- Expert in designing Parallel jobs using various stages like Join, Merge, Lookup, remove duplicates, Filter, Dataset, Lookup file set, Complex flat file, Modify, Aggregator, XML .
- Hands-on experience with Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SES, SQS, Lambda, EMR and other services of the AWS family .
- Created and configured new batch job in Denodo scheduler with email notification capabilities and Implemented Cluster setting for multiple Denodo node and created load balance for improving performance activity.
- Instantiated, created, and maintained CI/CD (continuous integration & deployment) pipelines and apply automation to environments and applications.
- Worked on various automation tools like GIT, Terraform, Ansible. Experienced in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension)
- Experienced with JSON based RESTful web services, and XML/QML based SOAP web services and also worked on various applications using python integrated IDEs like Sublime Text and PyCharm .
- Efficient Cloud Engineer with years of experience assembling cloud infrastructure. Utilizes strong managerial skills by negotiating with vendors and coordinating tasks with other IT team members. Implements best practices to create cloud functions, applications and databases.

SKILLS

- **Big Data Technologies:** Hadoop, MapReduce, HDFS, Sqoop, PIG, Hive, HBase, Oozie, Flume, NiFi, Kafka, Zookeeper, Yarn, Apache Spark, Mahout, Sparklib
- **Databases:** Oracle, MySQL, SQL Server, MongoDB, Cassandra, DynamoDB, PostgreSQL, Teradata, Cosmos.
- **Programming:** Python, PySpark, Scala, Java, C, C++, Shell script, Perl script, SQL
- **Cloud Technologies:** AWS, Microsoft Azure
- **Frameworks:** Django REST framework, MVC, Hortonworks
- **Tools:** PyCharm, Eclipse, Visual Studio, SQL*Plus, SQL Developer, TOAD, SQL Navigator, Query Analyzer, SQL Server Management Studio, SQL Assistance, Eclipse, Postman
- **Versioning tools:** SVN, Git, GitHub
- **Operating Systems:** Windows 7/8/XP/2008/2012, Ubuntu Linux, MacOS
- **Network Security:** Kerberos
- **Database Modelling:** Dimension Modeling, ER Modeling, Star Schema Modeling, Snowflake Modeling
- **Monitoring Tool:** Apache Airflow
- **Visualization/ Reporting:** Tableau, ggplot2, matplotlib, SSRS and Power BI
- **Machine Learning Techniques:** Linear & Logistic Regression, Classification and Regression Trees, Random Forest, Associative rules, NLP and Clustering.

EXPERIENCE

AWS DATA ENGINEER
01/2022 to 02/2022

Deloitte | Gilbert, AZ

- Designed and setup Enterprise Data Lake to provide support for various uses cases including Analytics, processing, storing and Reporting of voluminous, rapidly changing data
- Responsible for maintaining quality reference data in source by performing operations such as cleaning, transformation and ensuring Integrity in a relational environment by working closely with the stakeholders & solution architect
- Designed and developed Security Framework to provide fine grained access to objects in AWS S3 using AWS Lambda, DynamoDB
- Set up and worked on Kerberos authentication principals to establish secure network communication on cluster and testing of HDFS, Hive, Pig and MapReduce to access cluster for new users
- Performed end- to-end Architecture & implementation assessment of various AWS services like Amazon EMR, Redshift, S3
- Implemented the machine learning algorithms using python to predict the quantity a user might want to order for a specific item so we can automatically suggest using kinesis firehose and S3 data lake
- Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB
- Used Spark SQL for Scala & amp, Python interface that automatically converts RDD case classes to schema RDD
- Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response
- Creating Lambda functions with Boto3 to deregister unused AMIs in all application regions to reduce the cost for EC2 resources
- Importing & exporting database using SQL Server Integrations Services (SSIS) and Data Transformation Services (DTS Packages)
- Coded Teradata BTEQ scripts to load, transform data, fix defects like SCD 2 date chaining, cleaning up duplicates
- Developed reusable framework to be leveraged for future migrations that automates ETL from RDBMS systems to the Data Lake utilizing Spark Data Sources and Hive data objects
- Conducted Data blending, Data preparation using Alteryx and SQL for Tableau consumption and publishing data sources to Tableau server
- Developed Kibana Dashboards based on the Log stash data and Integrated different source and target systems into Elasticsearch for near real time log analysis of monitoring End to End transactions
- Implemented AWS Step Functions to automate and orchestrate the Amazon SageMaker related tasks such as publishing data to S3, training ML model and deploying it for prediction
- Integrated Apache Airflow with AWS to monitor multi-stage ML workflows with the tasks running on Amazon SageMaker
- Environment: AWS EMR, S3, RDS, Redshift, Lambda, Boto3, DynamoDB, Amazon SageMaker, Apache Spark, HBase, Apache Kafka, HIVE, SQOOP, Map Reduce, Snowflake, Apache Pig, Python, SSRS, Tableau
- Assessed organization technology infrastructure and managed cloud migration process.
- Configured computing, networking and security systems within cloud environment.
- Implemented cloud policies, managed technology requests and maintained service availability.

DATA ENGINEER
01/2016 to 11/2019

Cognizant Technology Solutions | Hatboro, PA

- Worked on Azure Data Factory to integrate data of both on-prem (MY SQL, Cassandra) and cloud (Blob storage, Azure SQL DB) and applied transformations to load back to Azure Synapse
- Managed, Configured and scheduled resources across the cluster using Azure Kubernetes Service
- Monitored Spark cluster using Log Analytics and Ambari Web UI
- Transitioned log storage from Cassandra to Azure SQL Datawarehouse and improved the query performance
- Involved in developing data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL
- Also Worked with Cosmos DB (SQL API and Mongo API)
- Develop dashboards and visualizations to help business users analyze data as well as providing data insight to upper management with a focus on Microsoft products like SQL Server Reporting Services (SSRS) and Power BI
- Performed the migration of large data sets to Databricks (Spark), create and administer cluster, load data, configure data pipelines, loading data from ADLS Gen2 to Databricks using ADF pipelines
- Created various pipelines to load the data from Azure data lake into Staging SQLDB and followed by to Azure SQL DB
- Created Databricks notebooks to streamline and curate the data for various business use cases and also mounted blob storage on Databrick
- Utilized Azure Logic Apps to build workflows to schedule and automate batch jobs by integrating apps, ADF pipelines, and other services like HTTP requests, email triggers etc
- Worked extensively on Azure data factory including data transformations, integration Runtimes, Azure Key Vaults, Triggers and migrating data factory pipelines to higher environments using ARM Templates
- Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Data bricks
- Environment: Azure SQL DW, Databrick, Azure Synapse, Cosmos DB, ADF, SSRS, Power BI, Azure Data lake, ARM, Azure HDInsight, Blob storage, Apache Spark.
- Adept in troubleshooting and identifying current issues and providing effective solutions.
- Managed performance monitoring and tuning while identifying and repairing issues within database realm.
- Identified key use cases and associated reference architectures for market segments and industry verticals.
- Designed surveys, opinion polls, and assessment tools to collect data.
- Tested, validated and reformulated models to foster accurate prediction of outcomes.
- Created graphs and charts detailing data analysis results.
- Recommended data analysis tools to address business issues.
- Developed new functions and applications to conduct analyses.
- Cleaned and manipulated raw data.
- Collaborated with solution architects to define database and analytics engagement strategies for operational territories and key accounts.

BIG DATA ENGINEER / HADOOP DEVELOPER
10/2013 to 12/2015

Novogradac & Co. Llp | Long Beach, CA

AnsibleDenodoDenodoCloudWatchAvroPySparkPySparkPySparkMLlibDataframeNiFiNiFi

- Interacted with business partners, Business Analysts and product owner to understand requirements and build scalable distributed data solutions using Hadoop ecosystem
- Developed Spark Streaming programs to process near real time data from Kafka, and process data with both stateless and state full transformations
- Worked with HIVE data warehouse infrastructure-creating tables, data distribution by implementing partitioning and bucketing, writing and optimizing the HQL queries
- Built and implemented automated procedures to split large files into smaller batches of data to facilitate FTP transfer which reduced 60% of execution time
- Worked on developing ETL processes (Data Stage Open Studio) to load data from multiple data sources to HDFS using FLUME and SQOOP, and performed structural modifications using Map Reduce, HIVE
- D eveloping Spark scripts, UDFS using both Spark DSL and Spark SQL query for data aggregation, querying, and writing data back into RDBMS through Sqoop
- Written multiple MapReduce Jobs using Java API, Pig and Hive for data extraction, transformation and aggregationAvrom multiple file formats including Parquet, Avro, XML, JSON, CSV, ORCFILE and other compressed file formats Codecs like gZip, Snappy, Lzo
- Strong understanding of Partitioning, bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance
- Developed PIG UDFs for manipulating the data according to Business Requirements and also worked on developing custom PIG Loaders
- Developing ETL pipelines in and out of data warehouse using combination of Python and Snowflakes SnowSQL, Writing SQL queries against Snowflake
- Experience in report writing using SQL Server Reporting Services (SSRS) and creating various types of reports like drill down, Parameterized, Cascading, Conditional, Table, Matrix, Chart and Sub Reports
- Used DataStax Spark connector which is used to store the data into Cassandra database or get the data from Cassandra database
- Wrote oozie scripts and setting up workflow using Apache Oozie workflow engine for managing and scheduling Hadoop jobs
- Worked on implementation of a log producer in Scala that watches for application logs, transform incremental log and sends them to a Kafka and Zookeeper based log collection platform
- Used Hive to analyze data ingested into HBase by using Hive-HBase integration and compute various metrics for reporting on the dashboard
- Transformed thPySpark using AWS Glue dynamic frames with PySpark; cataloged the transformed the data using Crawlers and scheduled the job and crawler using workflow feature
- Worked on installing cluster, commissioning & decommissioning of data node, name node recovery, capacity planning, and slots configuration
- Developed data pipeline programs with Spark Scala APIs, data aggregations with Hive, and formatting data (JSON) for visualization, and generatePySpark
- Environment: AWS, Cassandra, PySpark, Apache Spark, HBase, Apache Kafka, HIVE, SQOOP, FLUME, Apache oozie, Zookeeper, ETL, UDF, Map Reduce, Snowflake, Apache Pig, Python, Java, SSRS
- Onfidential
- Developed and implemented Hadoop code while observing coding standards.
- Optimized and tuned Hadoop environments and modified hardware to meet prescribed performance thresholds.
- Developed new functions and applications to conduct analyses.
- Created graphs and charts detailing data analysis results.
- Tested, validated and reformulated models to foster accurate prediction of outcomes.

PYTHON DEVELOPER
09/2012 to 10/2013

Fiserv | City, STATE

- AWS, S3, EC2, LAMBDA, EBS, IAM, Datadog, CloudTrail, CLI, Ansible, MySQL, Python, Git, Jenkins, DynamoDB, Cloud Watch, Docker, Kubernetes
- Leveraged open communication, collective decision-making and thorough reviews to create performant and scalable systems.
- Worked with server-side and front-end technologies and leveraged common design patterns to code dynamic and user-friendly systems.
- Implemented new API routes, architected new ORM structures and refactored code to boost application performance.
- Harnessed version control tools to coordinate project development and individual code submissions.
- Introduced cloud-based technologies into Python development to expand on-premise deployment options.
- Wrote clear and clean code for use in projects.
- Resolved customer issues by establishing workarounds and solutions to debug and create defect fixes.

EDUCATION AND TRAINING

Post Graduate | Data Engineering *02/2022*
Purdue University, West Lafayette, IN

Pc□t Graduate | Data Science And Business Analytics *09/2021*
University of Texas At Austin, Austin, TX

Ba□helor of Arts | Business Administration And Management *12/2009*
California State University , Fullerton, CA