**DATA ENGINEER**

Name:Kirthy

**Phone: (913)945-1166 Email:manoharv@softworldusa.com**

**Summary:**

* 7+ years of IT experience in all phases of Software Development with Agile methodology which includes User Interaction, Business Analysis/Modeling, Design & Development, Integration, Planning and testing, migration and documentation in applications using ETL pipelines and distributed applications.
* In - depth understanding of Spark Architecture including **Spark Core**, **Spark SQL**, **NO SQL, Postgres SQL,Data Frames, Snowflake, Spark Streaming – (Kafka), Kubernets, Structure streaming.**
* Extensively used ETL methodologies for supporting data extraction, transformations and loading processing in a corporate-side-ETL Solution using **Ab Initio/Informatica Power center 6.2/6.1/5.1.**
* Hands on experience on Google Cloud Platform (GCP) in all the bigdata products BigQuery, Cloud Data Proc, Google Cloud Storage, Composer (Air Flow as a service)
* Hands on experience in **SPARK SQL**, **DATAFRAMES**, **RDD** and worked on various file formats like text, CSV, **Parquet, JSON, AVRO, EXCEL** and loaded into **HIVE**
* Experience in building and architecting multiple Data pipelines, end to end ETL and ETL process for Data ingestion and transformation in GCP and coordinate task among the team.
* Experience in developing **Spark** with **Java** to compare the performance of SPARK with HIVE and SQL. Deploy Spark with **IBM Spectrum Conductor** as **cluster manager**.
* Experienced with Python frameworks like Webapp2 and, Flask.
* Highly experienced in ETL tool Ab Iniio in using GDE Designer, Very good working experience with all the Ab Initio components.
* In - depth understanding of **SnowFlake** cloud technology.
* In-Depth understanding of **SnowFlake Multi**-cluster Size and Credit Usage.
* Experience **with Snowflake** Multi-Cluster Warehouses.
* Exclusive work experience in converting **HIVE/SQL queries** into SPARK transformations using SPARK DataFrames, **Pandas**DataFrames, RDD’s, **Serialization, Deserialization** and loading into HDFS and/or GPFS
* Responsible for implementation of java integration components for JBPM and DROOLS
* Worked on Informatica Performance Tuning identifying and resolving performance bottlenecks in various like sources, targets, mappings and sessions.
* Experience with Design code, and debug operations, reporting, data analysis and web applications utilizing Python.
* Experience in GCP Dataproc, GCS Cloud functions, BigQuery, Azure Data Factory DataBricks.
* Experienced in Work load migration on AWS cloud
* Worked on Hadoop Distributed Files System (**HDFS**), General Parallel File System (**GPFS**) integration and extensive work on SPARK DataFrames which involve **transformations** and **actions** on DataFrames and RDD’s
* Experience in using **Broadcast** variables, **Accumulator** variables, **Cache**, **Persist** (modifying **Storage Levels**) of both DataFrames and RDD’s.
* **HIVE** tables and views creation in both **Optimized Row Columnar (ORC)** file format and Parquet. **Design Wrappers** in Shell Scripting which can recursively load the historic data from Oracle into the Hortonworks. Packaging and Promoting code to higher environments using **BitBucket**
* Build the Logical and Physical data model for Build the Logical and Physical data model for **snowflake** as per the changes required.
* Define roles, privileges required to access different database objects.
* In-depth knowledge of **Snowflake** Database, Schema and Tablestructures.
* Define virtual warehouse sizing for **Snowflake** for different type of workloads.as per the changes required.
* Define roles, privileges required to access different database objects.
* In-depth knowledge of Build the Logical and Physical data model for snowflake as per the changes required.
* Define roles, privileges required to access different database objects.
* In-depth knowledge of SnowflakeDatabase, Schema and Tablestructures.
* Define virtual warehouse sizing for Snowflake for different type of workloads., Schema and Tablestructures.
* Solid experience in the Analysis, Design, Coding and Testing of Data Warehousing solutions using **Ab Initio** both **Batch and Continuous graphs** development.
* Very Strong experience writing API’s/ and Web Framework Django.
* Good knowledge and understanding of **AWS (Amazon Web Services)**, Elastic Map Reduce (EMR), **Storage S3**, **RDS**,**EC2** instances and Data Warehousing, **Lamda**.
* Work with the network team to do the DNS registration of the servers built in AWS.
* Experienced in **ETL Informatica & Database Upgrade Testing.**
* Experience in branching, monitoring, and tagging the version across the environments using tools like **Git**, **SVN**
* Strong analytical and problem-solving skills and the ability to follow through with projects from inception to completion
* Ability to work effectively in cross-functional team environments, excellent communication, and interpersonal skills.
* Performed configuration, deployment and support of cloud services including Amazon Web Services(AWS).

**Tools and Technologies:**

* SPARK:2.3.0 to 3.0.1, IDE: IntelliJ IDEA 2020.3.1, Jupyter Notebook, DataBricks | Programming Languages:, PL/SQL, Python, Unix Shell scripting
* ETL Tools: Ab Initio GDE (Current Version: 3.3.2.1), Informatica 10.2, Co>Operating System (Current Version: 3.3.2.1), Enterprise Meta Environment (EME)
* Databases and tools: Hadoop Distributed File System (HDFS), TeraData, HIVE, Oracle 9i/10g/11g (current), DB2, Oracle SQL Developer (Current Version: 3.2.20.09) and Toad 7.x, SQL Loader
* Issue Tracking Tool: Bugzilla, ALM 12 - (HP Quality Center), Service Now
* Job Scheduling Tools: Conduct>It Plans, Autosys, Control M, Tivoli Workload Scheduler
* Mainframe: COBOL, JCL, Librarian, Platinum, ISPF, QMF and SPUFI
* Operating Systems: HP-UX, Windows, Sun Solaris 5.8, Linux 8x, MVS OS/390, z/OS
* Cloud technologies: Aws, EC2. S3, EMR

**Professional Experience**

**CLIENT: VISA, Austin, TX.**

 **Nov 2021 - Present**

**DATA ENGINEER**

**Responsibilities:**

* Building pyspark modules on IBM Spectrum
* Configured SPARK-SUBMIT along with option of SPECULATIVE EXECUTION to run jobs on cluster
* Load data into SPARK data frame and do in memory data computation to generate the output response.
* Created roles and access level privileges
* Experienced in writing live Real-time Processing and core jobs using Spark Streaming with Kafka as a data pipe-line system
* Involved in building database Model, APIs and Views utilizing Python, in order to build an interactive web based solution.
* Deployment automation, AWS EC2 instance mirroring, WebLogic domain creations and several proprietary middleware installations.
* Develop multi cloud strategies in better using GCP (for its PAAS) and Azure (for its SAAS).
* Created **Databricks** notebooks using SQL, and automated notebooks using jobs.
* Deliver a highly resilient, highly availability environment that is secure and regulatory-compliant. Ingest the data into Hortonworks and Hadoop distributed File Systems (HDFS) using Abinitio as well as Spark ingestion
* Created User Controls and simple animations using Java Script and Python.
* Developed ETL programs using Informatica Power center 9.6.1/9.5.1 to implement the business requirements.
* Working in the team as JBoss-Rules **Drools** BPM Specialist / Java Developer in a corporation wide modernization initiative for loans approval.
* Migrating rules from ‘Mind-Box’ to **Drools** (Red-Hat JBoss BRMS) with eclipse as a mediator. Solely in charge of provide and present of a more simple and efficient design and implementation of BPM rules engine technology to handle the complexity of the new Pricing. Ticketing initiatives by using Daptiv PPM and JIRA
* Involved in End to End migration of 800+ Object with 4TB Size from Sql server to Snowflake.
* Data moved from Sql Server Azure **snowflake** internal stage **Snowflake** with copy options.
* Created roles and access level privileges and taken care of **Snowflake** Admin Activity end to end.
* Schedule Alteryx set schedule location, add gallery, batch macros & Iterative macro.
* Have implemented chef provisioning for creation of on demand instances and instance infrastructure using the chef-client and bootstrapping using plugins against AWS EC2.
* Responsible for cleansing the data from source systems using **Ab Initio components s** such as Join, Dedup Sorted, Denormalize, Normalize, Reformat, Filter-by-Expression, Rollup.
* Integrated the **Drools** and JBPM to application framework, which involved dynamic creation of knowledgebase and Knowledge session.
* Designed, Configured, and Deployed **Amazon Web Services** (AWS) by utilizing the AWS stacks like **EC2, S3** focusing on high-availability, **Redshift Spectrum**, fault tolerance, and auto-scaling and worked for **Elastic Load Balancing** service on **AWS.**
* Designed Python batch processors to consume and produce various feeds.
* Developed Preprocessing job using Spark Data frames to flatten Json documents to flat file. Load D-Stream data into Spark RDD and do in memory data Computation to generate Output response.
* Used Informatica file watch events to pole the FTP sites for the external files.

**Environment:** SPARK, Java, AWS, Teradata, Python, Informatica, Alteryx, GCP, Alteryx, Ab Initio, JSON, HDFS, SQL scripting and Linux Shell scripting, Abinitio, parquet, avro. AWS.

**CLIENT:RBFCU, San Antonio, TX.**

**Dec 2020 – Oct 2021**

**Sr. Informatica Developer**

**Responsibilities:**

* Write Python scripts to parse JSON documents and load the data in database.
* Develop complex graphs using various Ab Initio components like Join, Reformat, Rollup, Scan, partition & de-partition components and schedule those jobs using Autosys.
* Developed data warehouse model in **snowflake** for over 100 datasets using whereScape
* Some Alteryx Admin Work, District, Workflows, Collection, Private Studio, Controller & Server Set Up/
* Got involved in migrating on prem Hadoop system to using GCP(Google Cloud Platform).
* Wrote Python scripts to parse XML documents and load the data in database.
* Developed Preprocessing job using Spark Data frames to flatten Json documents to flat file. Load D-Stream data into Spark RDD and do in memory data Computation to generate Output response.
* Used **Drools** to handle complex business rules using Rule Sets and Rule Flows. Integrated the
* **Drools** and **JBPM** to application framework, which involved dynamic creation of knowledgebase and Knowledge session.
* Creating Reports in Looker based on Snowflake Connections. Snowflake SQL Writing SQL queries against Snowflake Developing scripts Unix, Python etc to do Extract, Load and Transform data 3+ years of experience in Data management (e.g. DW/BI) solutions.
* Validation of Looker report with Redshift database
* Perform all Linux operating system, disk management and patch management configurations, on Linux instances in AWS. Consuming the Web Services and integrating them in the applications.
* Unit testing & System Integration in DEV and SIT and also involved in UAT migration and setup to deploy the jobs in Autosys (by JIL scripts)
* Used Couchbase Python SDK to build applications that use Couchbase Server.
* Migrated previously written cron jobs to airflow/composer in GCP.

**Environment:**Ab Initio 3.3.2.1, Co>Operating System 3.3.2.2, Python, Oracle SQL Developer Version 3.2.20.09, Spark, AWS, GCP, Autosys: CA Workload Control Center, Alteryx, Tectia(R) Version 6.2.1.168, Optional: PuTTY - Release 0.54, Microsoft Windows Azure

**CLIENT: Kensium Solutions Pvt. Ltd , Illinois, US.**

 **DEC 2018 – Nov 2020**

**DATA Warehouse Developer**

**Responsibilities:**

* Developed high level and low-level design document for processing each feed and documenting the various implementation done during the course of the project.
* Implement and maintain the monitoring and alerting of production and corporate servers/storage using Cloud Watch. Migrated applications from internal data center to AWS
* Used Spark Streaming APIs to perform transformations and actions on the fly for building common learner data model which gets the data from Kafka in Near real time and persist it to Cassandra.
* Closely interacted with the data modeling team and mappers regarding the mapping document to generate mappings from legacy source system to target database fields.
* Leveraged cloud and GPU computing technologies for automated machine learning and analytics pipelines, such as AWS, GCP.
* Developed Complex Ab Initio graphs, XFR’s to derive new fields that solve various business requirements by extensive usage of lookup’s, lookup local, In-Memory Joins and rollups to speed up various Ab Initio Graphs.
* Maintained the user accounts RDB, Dynamo DB, SES, SQS and SNS services in AWS cloud. Created and managed a Docker deployment pipeline for custom application images in the cloud using Jenkins
* Performed data validation before moving the data into staging areas using in built functions like is valid, first defined, string substring, is blank, string length and string index.
* Analyzing various logs that are been generating and predicting/forecasting next occurrence of event with various Python
* Worked with the Spark for improving performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Spark MLlib, Data Frame, Pair RDD's, Spark YARN

**Environment**: Ab Initio GDE Version 3.0.6, Co>Operating System 3.0.6, Oracle SQL Developer, SPUFI, QMF, Control-M, Python, COBOL, JCL

**CLIENT: Eclature Technologies, India.**

 **Feb 2017- Nov 2018**

**ETL Developer**

**Responsibilities:**

* Experienced in loading the raw data into SparkData Frames and validate the data.
* Experienced in converting the validated RDDs into Data frames for further processing
* Was involved in setting up of apache airflow service in GCP.
* Implemented the Spark SQL code logic to join multiple data frames to generate application specific aggregated results.
* Experienced in fine tuning the jobs for better performance in the production cluster space.
* Extensively worked on distributed File Systems like Hadoop (HDFS) and Distributed Processing
* Worked with Log4j framework for logging debug, info & error data.
* Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDDs.
* Built various graphs for business decision making using Python mat plotlib library.
* Experience in using Avro, Parquet and JSON file formats, developed UDFs in Hive.
* Designed appropriate partitioning/bucketing schema to allow faster data retrieval during analysis.
* Utilize windows physical server with support, maintenance & troubleshooting on which Alteryx resides.
* Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.
* Experience working with various File-Formats (Parquet, Avro, dat& JSON) and Compressions (Snappy &Gzip).
* Worked in managing VMs in Amazon using AWS and EC2.
* Responsible for generating actionable insights from complex data to drive real business results for various applications teams and worked in Agile Methodology projects extensively.
* Worked with SCRUM team in delivering agreed user stories on time for every Sprint.
* Scraping website using Python Beautfile Soup, and then parsed it with XML.
* Implemented UNIX scripts to define the use case workflow and to process the data files and automate the jobs.

**Environment**: Spark, AWS, Java, HDFS, Hive, Shell scripting, Jenkins, Python, Intelli J, Git.

**CLIENT: Now Floats Technologies Ltd. India**

**July 2015- Jan 2017**

**ETL Developer**

**Responsibilities:**

* Design physical solution, test strategy and migration schedule from QA to PROD. Involve in all phases of testing including UAT, Regression and User Acceptance.
* Set up Alteryx server & managed server/configurations.
* Tune performance of Informatica session for large data files by increasing block size, data cache size, sequence buffer length and target based commit interval.
* All the jobs are integrated using complex Meanings and Workflows using Informatica power center designer and workflow manager.
* Created Complex mappings using Unconnected, Lookup, and Aggregate and Router transformations for populating target table in efficient manner.
* Implemented & followed organisation level **Informatica best Standards & Procedures.**
* Create ETL between different data warehousing such as snowflake & Redshift via Alteryx workflow.
* Created Mapplet and used them in different Mappings. Used sorter transformation and newly changed dynamic lookup Created events and tasks in the work flows using workflow manager
* Used Python library Beautiful Soup for web scrapping to extract data for building graphs.
* Developed Informatica mappings and also tuned them for better performance with PL/SQL Procedures/Functions to build business rules to load data.
* Created Schema objects like Indexes, Views, and Sequences.
* Designed and Developed Oracle PL/SQL and UNIX Shell Scripts, Data Import/Export.
* Developed UNIX shell scripts for running batch jobs and scheduling them