Vishal Srivastava

Mobile: +91-9971153380 Email: srivis20@gmail.com

Headline

AWS and BigData Developer. Working as Senior Data Engineer. Having experience in Designing/Developing end to end automated, scalable ETL products from scratch and Data warehouse/Data Lake/Data modelling solutions in multiple projects. Implemented automated CI/CD pipeline.

Professional Experience

- Vishal is having 8 years of experience in IT Industry and working as Senior Consultant with Deloitte Consulting LLP.
- He has worked with clients, stake holders directly to gather the requirements in multiple projects.
- He is working as Technical Lead and involved in proposals, estimations, designing etc.
- **He has experience in designing and developing ETL products.**
- He has implemented end to end automated and large-scale data ingestion pipeline in multiple projects in different technologies like **Python, PySpark, Hive, Impala, RedShift, Athena, Aurora, PostgreSQL, Lambda, Kinesis, Step Function and various AWS Technologies.**
- **He has developed on ETL involving multiple layers of transformations.**
- **He has containerized ETL pipeline using Docker, ECS, ECR and Fargate.**
- **He is strong knowledge in writing and understanding complex SQL.**
- **He has also implemented GxP (General Practice) compliant ETL data pipeline which captures lineage of data extensively.**
- He is experienced with large volume data warehouse, data lakes and data modelling solutions.
- **Hands on experience in structured and semi-structured data.**
- He also knows AWS infrastructure services.
- **He has also implemented automated CI/CD pipeline using Jenkins.**
- Developing **analytical reports** by analyzing, parsing and extracting relevant information from **log data.**
- He has experience with industry standard data models **OMOP and FHIR(HL7).**
- Automating deployment with CloudFormation.
- He is able to perform POCs within limited time.
- Exploring and optimizing the framework for better client experience.
- He has led the team from offshore.
- He has worked on agile methodology. Also, involved in planning the Development/SIT/Deployment strategies.

Work Experience

Senior Consultant at Deloitte Consulting LLP
Apr 2016 – Present

System Engineer at IBM India Pvt. Ltd.
Mar 2013 – Apr 2016

Skill Set

Technologies/Languages: PySpark, Impala, Hive, SQL, RedShift, AWS Lambda, S3, EMR, EC2, SNS, Docker, Python, CloudFormation, ECS, ECR, Athena, Kinesis, Shell Scripting, Step FUnctions
Tools: PyCharm
OS: Linux

Education

Post Graduate Diploma in Advance Computing (A Grade - 2013)
Center for Development of Advance Computing (CDAC-ACTS), Pune

B.Tech. (73.4% - 2011)
Gautam Buddha Technical University, Lucknow

Intermediate (73.4% - 2007)
C.N.S. Inter College, Kanpur

High School (76.5% - 2005)
C.N.S. Inter College, Kanpur

Projects

- **Gilead**

Building analytical platform for data scientists. Data received from various sources in different formats (csv, json, fixed width). As part of this project a rule based data pipeline (ETL) has been developed to ingest the data into RedShift tables and S3 and Athena can be created on top. Pipeline has rich auditing features as well. This pipeline provides RWE (Real world data evidence) platform to the data scientists. Also, another pipeline has been developed for standard SDTM format The data pipeline is completely automated and there is no manual intervention.

Roles & Responsibility:

- Worked as ETL lead in the project. Requirement gathering, planning, designing, estimations and team handling.
- Designing, Data modelling, Developing and managing data pipeline based on **ECS, Athena, Redshift, PySpark, json and delimited.**
- Developing complex transformation logic like truncate load, incremental, upsert, handling separate header files, custom error handling.
- Developing extensive audit framework.

- **Broadcom (MyPathToWork: Contact Tracing Project)**

In this implementation a solution has been developed for post Covid workforce management for an organization. This solution is based on FHIR data model. User data comes in FHIR standard json format which is required for analytical platform for contact tracing. Another aspect in the solution was to analyze wifi log data to prepare a analytical sanitization report of Covid +ve person's place. His role is to develop an ETL pipeline to normalize complex json data and make it ready for analytical tools like Athena. Also, generating report of users who are eligible or not eligible for office. Developed ETL pipeline to read, analyze and parse log data and generate the report.

Roles & Responsibility:

- Designing, developing ETL data pipeline based on **Lambda + Athena + Aurora DB + Kinesis** to handle json data and normalize it to load in Athena.
- **Analyzing wifi log data and parsing it to develop analytical report**
- **Developing complex transformation logic on incremental, upsert type loading with SQL.**
- Validating/cleaning/ingesting data into the system.

- **Merck (DevOps)**

  In this implementation a fully automated Jenkins pipeline has been developed. This solution includes an end to end process starting from code checkin to SCM till deployment to docker. The pipeline fetches data from Git then runs Sonar on it for coding standardization then builds to code and finally deployed to a docker image. Whenever required this image can be containerized to run and application will be available. Also, communication between two docker container was established for the independent backend and frontend application modules.

Roles & Responsibility:

- **Designing, Developing and managing Jenkins pipeline based on docker.**
- **Communication between two docker containers.**
- Regularly communicating with clients, preparing demos.
- Leading the whole Jenkins team.

- **Research Trust**

  This is a Deloitte product based on health care system. The model is based on FHIR standards. This data model captures very vast variety of patient related data. It captures all the instances that patient encounters during his cycle like diagnosis, testing, procedures etc. This data is then further used by analytical tools in preparing cohorts, analyzing patient journey etc.

Roles & Responsibility:

- Building and managing data pipeline based on **impala and shell and integrating hive ingestion with impala, Athena, Redshift.**
- Containerized ETL using **ECR**, **ECS and Fargate.**
- Writing complex SQL transformations.
- Working on ingestion framework and loading the data based.
- Developing extensive auditing and error handling like batch, job, file level auditing and record level error handling based on business and model logic.

- **Merck (RWE: Real World Evidence)**

Data is received from various sources in different formats (csv, json, fixed width). As part of this project a data pipeline (ETL) has been developed in pyspark to ingest the data into RedShift tables and on S3 so that RedShift Spectrum can be created on top. This pipeline provides RWE (Real world data evidence) platform to the data scientists. When data has been ingested data is converted to industry accepted RT data model. The data pipeline is completely automated and there is no manual intervention. The framework provides end to end lineage.

Roles & Responsibility:

- Worked as Technical lead.
- Designing, Developing and managing data pipeline based on **Lambda + PySpark + RedShift to handle json, delimited and fixed width datasets.**
- **Created an unzipping utility which will spin up ec2 on the fly, do the unzipping process and terminate ec2 instance.**

- Developing complex transformation logic like truncate load, incremental, upsert, handling separate header files, error record handling.

- **Celgene (RWE: Real World Evidence)**

It's an implementation of data pipeline in Hadoop environment. Data is received from various sources in different formats (delimited, fixed width). As part of the project a data pipeline framework on impala + shell scripts and building data pipeline to load data across cluster. Data cleaning/validating data/ingesting data are the major tasks. This pipeline provides RWE platform to data scientists for analysis. On top of the data Tableau visualizations are created for clients/data scientists for analysis.

Roles & Responsibility:

- Building and managing data pipeline based on **PySpark to handle json and delimited datasets.**
- Building and managing data pipeline based on **impala + shell and integrating hive ingestion with impala.**
- Owning RT ETL for multiple myeloma and designing data mapping according to model specification.

- **Kohl**

The objective of the project is to try different capabilities of hadoop and other open source tools and advice client with the best approach for their requirement.

Roles & Responsibility:

- Figure out the capabilities of the tools and provide solution based on feasibility.
- Integrating ElasticSearch and MongoDB for better searching and storing capability.
- Integrating Hive and HBase for better storing and analytic capability.
- Prepare a model on Neo4J to be specific, for better suggestion in online retail system.

- **Ecolab**

Worked as Application Support Analyst. Scope of the project includes enhancements and support. In this project the work is done to maintain the daily business objective. Along with this there are minor enhancement as well. Development from scratch is also involved. Ecolab works in distribution sector. Daily sales, purchase, inventory management is in the scope of the project. Apart from this customer management, employee management, their compensation all is managed.

Roles & Responsibility:

- Gathering and understanding clients' requirement. Understanding the core business process.
- Finding the recurring problem which has been continuing in the business system and resolving them with efficiency.
- Owning of biggest application of the project FAM, Compensation.
- Writing complex SQL statements for bulk data movements.
- Enhancing the process as and when required by business.

Personal Details

Address: A1-404, Unitech Uniworld Gardens II, Sector 47, Gurugram
DOB: 10th Aug, 1989
Nationality: Indian
Lang. Known: Hindi, English