Shashank Mishra

Data Engineer

Phone: +91 9648581661; **Email:** shashankmmishra@gmail.com

TECHNICAL SKILLS

- **Programming Languages:** Python, Java, Scala, C++, C, Linux Shell Scripting
- **Frameworks:** Django, Scala Play Framework, Android Studio
- **Web Development:** JavaScript, AJAX, JQuery, HTML, CSS, BootStrap, NGINX
- **Databases & Connectors:** MySQL, Vektorwise, ElasticSearch, Cassandra, HBase, Logstash, Kafka Connect
- **Distributed Systems Frameworks:** Hadoop, Spark, Hive, Kafka, BigQuery, DataRush
- **Dashboarding & MonitoringTools:** Grafana, Kibana, DataStudio, DataDog **,** QuickSight
- **Workflow Management:** Azkaban, Airflow
- **AWS Services:** S3, EMR, Lambda, Cloudwatch, DynamoDB **,** Redshift, Athena, Appflow, Glue, Secret Manager, SNS, SQS
- **Build Tools & Containers:** SBT, Maven, Jenkins, Docker
- **Control Systems and Documentation:** Git, BitBucket, SVN, Jira, Confluence

WORK EXPERIENCE

Data Engineer – III

Expedia **Nov 2021 – Present**

One checkout Platform – Real time streaming application for Financial data

Tech Stack – Flink, Python, Kafka, Oracle, Kafka Connect, Docker, AWS, Airflow

- Building **one checkout platform** for different business domains of Expedia i.e; Vrbo, Hotels, Car Rentals, Lodging
- **Real time** streaming platform captures & process financial data for **Accounting** and helps suppliers to track total amount to be paid via Auto Pay or Request Pay model
- Crafted a scalable streaming solution using **Flink StateFun** to handle ~**800k** booking streams each day

Data Engineer

Amazon **Mar 2020 – Nov 2021**

Salesforce to Redshift Ingestion - Migration from Informatica to Native AWS

Tech Stack – Salesforce, Informatica, S3, Lambda, Glue, AppFlow, Redshift, SNS

- Crafted generic **scalable Native AWS** solution for Salesforce to Redshift ingestion
- It helped to move ingestion pipelines from third party tool **Informatica** and **saved cost** for heavy license fee
- This generic framework helped other business units for smooth ingestion of newly onboarded Salesforce object into Redshift Datalake

Incremental Ingestion pipeline – Employee Benefits Data

Tech Stack – Shell Scripting, AWS CLI, S3, EMR, Glue, Redshift, SNS, QuickSight,PySpark

- Build generic & optimized ingestion pipeline for highly **critical** & **confidential** Employee Benefits Data
- Pipeline is designed in a way to handle **GB's of daily & weekly** data together for different use cases like Audit, Payroll, **Reimbursement**, Education Reimbursement etc
- Took complete ownership and worked closely with business teams to understand the requirements & deliver enriching **dashboards**

Pipeline Optimization & Enhancement

Tech Stack – Shell Scripting, AWS CLI, S3, EMR, Glue, Redshift, SNS, QuickSight, PySpark, Lambda

- Enhanced & optimized multiple pipelines, built for different business units like Peoplesoft, Audit, AEM, **Immigration**, **Accurate**, MyDocs, Background Verification etc
- Reduced execution time by **50%** and improved **alerting** system for different edge cases
- Leadership principals like **Customer Obsession**, Earn Trust and **Think Big**, helped me to keep improving existing systems

Data Engineer

Paytm

Jan 2019 – Dec 2019

Data Ingestion & Sync Process

Tech Stack – Python, Hive, ElasticSearch, Scala Play Framework, SBT, EMR, Lambda, DynamoDB, Azkaban, Jenkins

- Crafted **data-sync** logic by prioritizing datasets (High/Medium/Low tag) based upon criticality to meet SLO
- Built premption logic to prioritize highly critical datasets when multiple low priority sync processes are running
- Designed **Rest API** in data ingestion for retention of **GA** data in order to optimize cluster space
- Added **exception handling** scenarios in data sync logic to fix multiple bugs
- Fix for missing **PG** data from Kafka for **UMP** panel - Created a new pipeline to ingest missing data from HDFS to ElasticSearch in case of cluster failure

Near Real Time Data Pipeline – POC

Tech Stack – Java, Spark, Kafka, Datastax Cassandra, Datastax studio, Zookeeper, Maven

- Crafted a **Cassandra** based real time ingestion pipeline for **marketplace** data in order to help DWH team to reduce request load from production MySQL. The Objective was to shift business users from production, to overcome **data leaks** & **security** issues
- Interacted with different business users to know about their use cases, ingestion tables, PII data and built data models accordingly for faster **insertion/updation** of data
- Setup web interface **Datastax Studio** for users to query real time data from Cassandra using LDAP authentication

Dehleez - Report Scheduling Tool

Tech Stack – Python, JavaScript, Django, Azkaban, Docker, Hive, Ajax, Bootstrap, REST API, DataDog

- Enhanced Paytm's proprietary report scheduling tool which is used by business users working on data analysis where they can schedule their reports by writing HIVE/MySQL/Cassandra queries and report output in various formats
- **Diff Checker** - Admins can check the difference between queries before approving reports
- **Time slot picker** to schedule a report - User can see scheduled reports for next 4 hours from intended schedule time and can pick the slot accordingly
- **Dump report output into S3 bucket** - User can take dump of report output into AWS S3 bucket
- **Cassandra Connector** - User can schedule reports having Cassandra query panels in addition with HIVE/MySQL

Software Engineer II

Opera Solutions

Jan 2017 – Jan 2019

Procurement Spend Optimization (Pharmaceutical)

- Developed CXO-level insights engine to manage **USD 60Bn**; engine enabled cost optimization using smart categorisation, benchmarking and **anomaly detection**
- Crafted a Big Data based solution; organised structured & unstructured data
- Built solution using Hadoop Ecosystem (HDFS, YARN), **Spark** and **Python**
- Built a **google translator API** based solution to automate legacy translation engine; improved record aggregation accuracy by 50% and saved team 120 hours/month

Trip Narrative Platform (Aviation)

- Deployed an end to end solution for a **leading US airlines**; Aggregated a 360 view of customer's engagement throughout the life-cycle of the trip
- Developed **data pipelines** from scratch; optimised data aggregation from 10+ independent sources and automated the ETL process to roll out the solution
- The solution powers a web application; used by **1000+ CSRs** and decision makers
- Built application on **RESTFUL API`s** using Hadoop Ecosystem (HDFS, YARN), **DataRush** Applications (Distributed Processing Engine), **SQL** and Python

EDUCATION

Master of Computer Applications

National Institute Of Technology Allahabad (MNNIT), U.P

Jul 2014 – May 2017, CPI 8.5/10 (NIMCET Rank 43 of 20k)

Bachelor In Computer Science

Lucknow University, U.P

Jun 2011 – Jun 2014, Per 70/100

RECENT ACHIEVEMENTS

- **Opera Ovation Award** : For exceptional work on a Trip Narrative Project, 2017
- **Runners Up** : Opera Python Hackathon; internal python coding competition, 2017
- Conducted a firm wide **Global Level** session for 200+ employees on SHM encompassing designing, executing and best practices on **workflows**
- Organised BigData session for India’s Analytics team of 50+; demonstrated usage of **HIVE** platform; personally conducted 10+ workshops
- **Geek Of The Month** : GeeksForGeeks; for contribution in technical content, 2016
- **Coordinator** : Computer Club Classes, NIT Allahabad, 2016
- **Dance Coordinator** : SWAGAT (Cultural festival) of NIT Allahabad, 2015
- **Volunteer, Pahal (NGO)** : conducted 100+ classes for 150+ slum kids, 2014
- **2 SILVER MEDALS** : HourRank 18 and Week of Code 30; Hackerrank 2017
- **3 BRONZE MEDALS** : Week of Code 29, HourRank 17 and World CodeSprint 9; Hackerank 2017
- **GLOBAL RANK 10** : Amongst 3 lac coders; solved over 35 coding challenges in 24 hours SPOJ, 2016
- **FINALIST** : REVENGE, 2015 Inter-college coding competition during Avishkar (Tech. Fest.), NIT Allahbad