|  |  |
| --- | --- |
| Prateek Dubey | |
| Profile:  Data Science Lead for ML Infrastructure team leading a team of Software Engineers and Machine Learning Engineers at Grab. I support DS product families like Ads & Data Insights, Marketplace, Transport, Geo and Deliveries by providing them a Machine Learning Platform on both Azure and AWS which is being used to build, train, deploy and manage models in production. We build solutions and services that bridge the gap between Data Science and Product. My work revolves around building, supporting and enhancing the existing ML Platforms and collaborate with product families to design Data Engineering pipelines, Feature Engineering, ML Model Serving pipelines and to provide Cloud based Architecture solutions. I have extensively worked on to carry out day to day work involving Cloud Engineering, Data Science/ Engineering, and MLOps using Python, Databricks, Kubernetes, Docker, Hadoop, Spark, Terraform, AWS and Azure. | |
| Personal details Nationality Indian Location Singapore Languages English Stanton House Contact Name Laura Taylor | Education & Qualifications BEng Computer Engineering (First Class / Distinction) 2009 – 2013: University of Pune, IND |

Employment History

GRAB (BU – Data Science), SG

Data Science Lead (AWS, AZURE, K8s, ML, Big Data)

Sep 2018 – To Date

* Technical lead for ML Infrastructure team and responsible for looking after Cloud Architectural design solutions over AWS and Azure for DS Product families. Involved in MLOps, Cloud, Data Engineering, and Data Science.
* Leading a team of 4 Engineers to build, manage and support our existing ML Platforms on AWS and Azure along with supporting teams in their ML projects by building jobs related to Feature Engineering, ML, DevOps, K8s etc.
* DS PIC for vendor engagement with AWS, Microsoft and Databricks and also owner for all our DS Cloud Platforms.
* Worked with GrabFood and Azure CAT team to develop GrabFood restaurant behavioral recommendation system on Azure thus increasing Click through Rate by 170% compared to existing recommendation systems in Grab.
* Streamlining the existing DS Infrastructure on AWS and making it more reliable, fault tolerant and resilient.
* Developed a multiuser JupyterHub Data Science ML Platform on AWS EKS for ML and Analytics.
* Worked on Kubeflow Pipelines to automate the ML pipelines on AWS EKS.
* Driving AI Adoption across the organization by pitching adoption of Databricks as a preferred ML Platform on Azure.
* Cloud platform automation using Terraform and Azure ARM. Using Helm to build K8s compatible services on EKS and AKS, and thus transitioning to a fully managed K8s platform for our production deployments on AWS. Using Spinnaker for deploying models in production on K8s.
* Lead the team to architect and develop multiple services like Spark UI Proxy, Jupyterhub resource tracking, AWS Subnet/ IPv4 tracking, Datadog alerts for EKS services, Agartha – Docker build Automation using Git Pull Requests.
* Automated AWS Key Rotation process in Data Science on AWS using Lambda, SNS, SES, CloudWatch
* Developed STS temporary token generation web service using Flask, Nginx and Python and deployed on EKS

Royal Bank of Scotland (BU – Enterprise Solutions), Gurgaon, IND

Software Engineer (Hadoop / Spark / AWS Developer)

Jul 2017 – Aug 2018

* Worked as part of Cloud Engineering Team and responsible for migrating Athena SAS Datawarehouse onto Public Cloud – AWS for Risk and Finance Solutions.
* Designed a public cloud data lake on AWS, which will be utilized by Data Engineering & Analytics BU for ETL workloads, data analytics and Risk modelling thus reducing the operational and infrastructural cost by ten folds.
* Build AWS infrastructure for Risk and Finance Solutions using Terraform.
* Designed and Developed a Python Parser to auto-convert HiveQL codes into equivalent PySpark (Spark SQL) jobs to leverage the Spark capabilities on AWS EMR, thus reducing conversion time by over 90%.
* Reduced workflow creation time by 80% through an automated Oozie workflow creation framework.
* Designed processes for seamless monitoring like process to publish all Active EMR Cluster and SageMaker instances running across all regions. Created using Python Boto3 library, running on AWS Lambda, pushing notifications to business users using AWS SES and scheduled using CloudWatch.
* Enabling Amazon SageMaker for Risk Modelling (Machine Learning) to leverage Spark on SageMaker both locally and through remote EMR Cluster. Persisted SageMaker Jupyter Notebooks in S3 instead of local EBS volume.
* Setup Apache Airflow to schedule our Core Spark ETL Workflows.
* Worked on a framework for orchestration and monitoring of our Core EMR Cluster using AWS Lambda, CloudWatch, SNS and CloudFormation.
* Training Retail Credit Risk colleagues on AWS services and Spark (using Python) and conducted some internal trainings for the same.

EXL (BU – Barclays Decision Analytics), Gurgaon, IND

Business Analyst (Hadoop / Spark Developer)

Dec 2016 – Jul 2017

* Worked as part of Decision Analytics team which was spearheading Big Data adoption for Barclays bank thus reducing operational costs and delivering better insights from data.
* Worked on migrating SAS Datawarehouse onto Big Data platform - Cloudera for Risk Solutions vertical of Barclays.
* Designed an on-premise data lake using Cloudera CDH, which will be utilized by Data Engineering & Analytics track for ETL workloads, data analytics and modelling.
* Designing and architecting the ETL data pipeline solutions to move data and ETL from SAS to Big Data platform using PySpark and HiveQL.
* Designing a near real-time data pipeline using Spark streaming, PySpark and HBase.
* Reduced metadata migration time from Teradata to Hadoop by 95% through an automation suite created using UNIX.
* Reduced historical data migration time from SAS to Hadoop by 70% through an automation suite created using UNIX.
* Training peers on Hadoop and Spark technology stack along with Unix & Shell scripting

AON (BU – Business Intelligence), Gurgaon, IND

Senior Software Engineer (Hadoop Developer(

Jan 2014 – Dec 2016

* Worked on multiple end to end Big Data project deliverables involving data ingestion, developing ETL, designing a data lake, data migration, creating unit test plans, proof of concepts etc.
* Subject matter expert for Legacy Data warehouse running on Unix and SQL platform using IBM Informix database.
* Designed and developed HRBPO Attendance Tracking System, which is used to track logging hours of CSR’s.
* Designed and developed an automation framework using Hive and Unix to catalog the PRE-and-POST CDH upgrade checks, thus reducing the testing time from 5 business working days to 1 hour.
* Carried out POC’s on varied Hadoop technologies like HDFS Snapshots, High Availability of HDFS/Hive/YARN/HUE, NOSQL database - HBASE, Active Directory implementation on Hadoop Servers, Hive on Spark, Scala on Spark etc.
* Supporting BI Infrastructure team with CDH upgrades and setting up QC cluster for Big Data platform.
* Responsible for production ETL workload job monitoring using Control-M, ETL fixes and ETL enhancements raised by business stakeholders and users. Designed and developed ETL codes in PIG Latin, HiveQL, Unix and SQL.
* Expertise in working with various file formats like JSON, PARQUET, TEXT, XML, RCFILE, ORC.
* Additional responsibilities include coordinating with On-Shore business stakeholders on weekly calls for status/project updates, training junior peers, triaging of issues through Maestro etc

Projects

ML PLAT FORMS (GR AB )

* I lead a team of MLE’s and designed ML Platforms at Grab for DS Product Families on both AWS and Azure Cloud. My team also support these platforms for all production workloads. On AWS we developed an in-house product called Libera which is a multi-user Jupyterhub environment that runs entirely on EKS and K8s acts as our Spark compute cluster. We use Spinnaker for Model Deployment on EKS and Kubeflow for Model Training. On Azure we decided to choose Databricks as the preferred notebook environment for ML needs at Grab and use MLFlow for Model Training.
* **Technologies Used**: Python, ML, K8s, Docker, Terraform, ARM, Databricks, AWS, Azure, Kubeflow, MLFLow

GRAB FOOD RESTAU RA NT RECOMMENDATION MODEL (GRAB )

* I worked with GrabFood product/tech family and Microsoft Azure CAT team Data Scientists to design a restaurant recommendation model based on customer past order history, so that we can refine our existing models and provide better recommendations related to eateries to our customers. The model reflected a 170% increase in CTR in production.
* **Technologies Used**: PySpark, ML, Databricks, MLFlow, Azure

ATHENA RE- PLATFO RM ING *–* SAS DATAWARE HOUSE MIGR ATION TO AWS ( RBS)

* I worked as part of Cloud Engineering team in RBS (Pilot team for AWS in RBS) and helped Risk and Finance solutions to move their legacy Datawarehouse platform from SAS to AWS EMR, Glue, Athena, Quicksight and SageMaker. All SAS codes were re-written in PySpark and scheduled via Airflow on EMR Spark Clusters.
* **Technologies Used**: Hadoop, Spark, AWS, Python, SAS, Airflow, Unix, Shell Scripting

BA RCLAYS BU REAU *–* SAS DATAW AREHOUSE MIGR ATION TO CLOUD ERA ( EXL *–* CLIENT BA RCLAYS )

* I worked as part of EXL Barclays Decision Analytics team and helped Risk Solutions to move their legacy DWH platform from SAS to Cloudera Big Data Platform. All SAS codes were re-written in PySpark and scheduled via Oozie on Cloudera.
* **Technologies Used**: PySpark, Unix, Shell Scripting, HiveQL, Teradata, SAS, Impala, SQOOP, Python, HBase, SPARK

INFORM IX SHUTDOWN *–* LEGA CY DATAWAREH OUSE MIGR ATION TO CLOUDERA ( AON)

* I worked as a Hadoop Developer in LDW Migration team wherein we migrated over 2000 codes written in Unix & SQL by rewriting them in Pig Latin and porting them over to Cloudera Big Data Platform.
* **Technologies Used**: Unix, SQL, Shell Scripting, PIG Latin, Beeline, Hive, Impala, SQOOP, CONTROL-M, Informix

Systems Skills

* Hadoop Stack: Cloudera CDH, Hadoop, Pig Latin, Impala, HiveQL, Hive, Beeline, Sqoop, Flume, HUE, YARN
* Spark Stack: Spark SQL, Dataframes, RDD, PySpark, Spark on Mesos and Kubernetes, Delta Lake, Parquet
* AWS Stack: EKS, ECR, EMR, EC2, Glue, Lambda, S3, DynamoDB, Athena, CloudWatch, AWS CLI, SageMaker, SES, SNS, SQS, IAM, Cloud9, VPC, Route53
* Azure Stack: AKS, ACR, VM, ADF, HD Insights, Azure Databricks, Blob, ADLS Gen2, Azure CLI, Azure AD, Key Vault
* Machine Learning Toolset: Kubeflow, MLFlow
* SAS: Base SAS, PROC SQL, Macros
* Programming & Scripting languages: Python, Scala, Unix, Shell Scripting
* Development Notebooks and IDE: Zeppelin, Jupyter, Databricks, Visual Studio, PyCharm & Spyder
* Workflow Scheduler: Control-M, Apache Oozie, Apache Airflow, Papermill
* Relational & NOSQL Databases: IBM DB2, Informix, MySQL, Teradata, HBase, Presto
* CI/CD and Artifactory: JIRA, Gitlab, Confluence, Wiki, Jenkins, Gitlab CI, Spinnaker
* Infrastructure as Code: Terraform, Azure ARM
* Container Orchestration: Docker, Kubernetes, Helm