Mengfan (Fox) SHAN

Address: 466 Tianbao Road, Shanghai, 200086 **Tel:** 18017661124 **Email:** dd.famous@gmail.com

WORKING EXPERIENCE

Shuhe Group, Shanghai Since April, 2019

I am now working in Shanghai Shuhe Group, in charge of Platform Team, BigData Department. My work is mainly related to BigData middle-wares, such as Hadoop, Hive, HBase, Spark, etc. I design, deploy, monitor and optimize all BigData services, by cooperating with Data Warehouse Teams and Data Application Teams to provide high-available data services for both customers and data scientists inside company. I also take responsibility of problem solving, new component introducing, etc, to improve performance and stability of BigData platforms.

Works Applications, Shanghai August, 2016 April, 2019

I joined Team EDP2 (Base Technologies Div) in 12/2016 and became the TL in 12/2017. EDP2 is a developing and operating platform for enterprise application on cloud, for higher speed, better stability and cost efficiency. EDP2 includes a server for job management and a Hadoop/Spark cluster for job execution.

PROJECTS

FeatureHub

FeatureHub is a platform managing and providing features, for training models and online-prediction. Unlike traditional feature platforms, which focus on data processing, FeatureHub provides processed data for both online and offline services with the unified calibre. It provides real time features for online services with low latency, as well as historical data for training.

Scalable

In original design, the data warehouse is on one single Hadoop based cluster. The cluster is slow to scale up and hard to scale down and it is also a SPoF for most services. HDFS is replaced by object storages and data processing jobs are separated to many state-less auto-scale EMR clusters for low cost and better availability.

Customization

Open source softwares could hardly cover all requirements. We did a number of customizations on open source softwares, for column based sensitive data check, SQL filter, data relationship. We also make AdHoc queries can join data from different sources, such as Hive, HBase, Mongo, Kudu, etc, and sometimes patch bugs such as HDFS-9406, HIVE-19994.

SLA

Ensure stability and performance of BigData services. Services provided to customers should have an SLA of 99.99% within 200ms. For example, we had a HBase cluster was unstable. The average response time is over 100ms and MTTR always took hours. After re-designing, the cluster never crush again, the average response time is on 20ms and the performance is extremely improved.

Multi-Clouds

The BigData cluster was migrated from Cloudera to AWS EMR, then from AWS to AliCloud. To minimize migration impact and downtime to business, the platform foundation needs to be compatible with changes and we need to provides all kinds of automation tools for migration and further management.

SKILLS

Languages: Java, python, Ansible, shell
Middleware: Hive, HBase, Hadoop, Spark, Flume, Flink, Zookeeper, Kafka, neo4j, Presto, HugeGraph
Skills: AWS, Aliyun

EDUCATIONS

Nanjing University, Nanjing 09/2010 - 06/2016
Ph.D. in Computer Software and Theory, Department of Computer Science and Technology
University of Minnesota, Twin Cities, MN 10/2014 - 10/2015
Visiting Scholar supervised by Prof. Tian He, Department of Computer Science and Engineering
Nanjing University, Nanjing 09/2006 - 06/2010
Bachelor, Department of Computer Science and Technology

REPOSITORIES

GitHub https://github.com/ddfamou