JIM GUAN/关键

- **Job Intension**: Related to **Big Data**, **Algorithm** or **Python Development**
- **Mobile**: 15624901008
- **Email**: guanwwjian@163.com

About Me

- With 2 years work experience as a Big Data Algorithm Engineer related to personalized push notification
- With wide-ranging professional skills
- Good at innovate

Education and Work Experience：

2017.4 ~	Sogou	Big Data Algorithm Engineer
2016.6 ~ 2016.12	SAP	Java Development Intern
2014.9 ~ 2017.4	Beijing Institute of Technology	Master of Computer Science and Technology
2010.9 ~ 2014.7	Tongji University	Bachelor of Computer Science and Technology

Professional Skills

- Primary programming languages: **Python, Linux Shell, SQL**
- IDE: **vim**
- Big data processing: **Spark** (Using in prediction process for personalized push notification), **Hadoop Streaming** (Using in log cleaning), **Hive** (Using in data statistics)
- Machine Learning: **TensorFlow** (Using in Wide & Deep CTR estimation Model), **Spark ML**
- Database: **Redis, MySQL, SQLite**
- Restful interface framework: **Sanic** (personalized push service)
- Restful interface service deployment tools: **nginx, PM2, expect**
- Other programming languages: **Java** (custom Hadoop InputFormat/OutputFormat), **C++** (personalized backend)
- Statistics and visualization: **Kylin, Redash**
- English skill: **CET-6**
- Understand the status and details of push technology in China

Work Experience

“ToptenToday APP” push notification process technical side person in charge

- Push notification introduction: Push notification is an important way for an APP to enhance user retention. For users who have installed APP, It’s possibly that the notification message will be displayed when the APP process is dead. Once the user clicks the notification bar. Message, the app will be activated to the foreground, guiding users to use the APP

- User scale:
  - DAU: 500,000
  - Count of push target users: 5,000,000
- Design the push Payload interface format in the client
- Design push-related Pingback format
- Push-related statistical work based on the pingback receipt log
- Collect and update client "pushid" based on pingback receipt log
- Coordinate the work of client, front end, back end and operational staff
- Participate in the work of push channel research and push service architecture design

"ToptenToday APP" CTR Estimation Model for Personalized Push notification

- Function: Calculate the CTR estimated score for the user-article pair
- Using programming languages: **Python, Shell**
- Using the **Wide&Deep** model built on **Tensorflow**
- Using **Spark** to distributed preprocess the training data, extract feature and process features into **TFRecord** format
- Training model with GPU and dynamic learning rate
- Calculate the CTR estimated score by loading tensorflow model in the Spark node, and there is no need to modify the code of prediction process when the Tensorflow model structure changes

"ToptenToday APP" Personalized Push Service

- Function: After receiving the push command, push the article with the highest CTR estimated score for each user, aggregate the users according to the article, and submit the push user collection of each article to the push service.
- Using programming languages: **Python3**
- Workflow: Maintain a priority queue for each user, receive the article score calculated by the CTR estimation model, and insert the article score into the priority queue of the corresponding user. The article with the highest score is at the head of the priority queue. Whenever a push command is received, every priority queue's head article is poped, aggregated by article and submitted for push.
- Using **Redis**'s **Sorted Set** structure as priority queue
- Restful interface service using **Sanic** framework
- In order to guarantee the push speed, the producer-consumer pattern provided by the **Celery** framework is used to get the top article of each user in a distributed manner, so that the sending speed reaches 1 million users/minute.

"ToptenToday APP" general statistical process based on pingback receipt log

- Function: According to the configuration, perform minute-level PV and day-level UV statistics on the "ToptenToday APP" pingback receipt log. This process supports filtering and classifying logs according to multiple different field dimensions.
- Using programming languages: **Python, Shell, SQL**
- Advantage:
  - Simple configuration: only one line of configuration items is required for each statistical dimension combination in the configuration file
  - The configuration item does not need to be modified frequently: configuration does not need to be modified when there is new value of

existing field in the pingback receipt log. The related statistics of the new value will be listed in the statistics table automatically.
    - Simple maintenance: When new fields appear in the log, modify the hive view without creating a new Hive table, and new fields will be supported
- Architecture:
    - **Hadoop Streaming** cleans the data
    - Load cleaned data with Hive
    - Generate SQL statements based on the configuration file and submit to **Hive**
    - Insert Hive statistics into **Mysql**
    - Visualize the statistics in Mysql using **Redash**

Maintenance text search personalized recomend process named "Guesswant" in "Sogou Search APP"

- Introduction: Responsible for maintaining the existing "Guesswant" service, which is to recommend personalized search query to enhance the user experience.
- Using programming languages: **C++, Python, Shell**
- The offline process uses the **CF**(collaborative filtering) algorithm to calculate the "Search History - Recommended Words mapping", which is based on the user's search history in the pingback receipt log.
- Using Redis to store the "Search History - Recommended Words mapping"
- The online Restful interface is developed in C++. It is responsible for receiving user's search history, and returning the corresponding recommendation search query according to "Search History - Recommended Words mapping".
- Using the **AC automaton** algorithm as blacklist fuzzy matching method