**Tell me a little about yourself**

My name is Zheyi and you can call me Chris if you like. I have two master degrees, one is in the University of Bristol with an double EE degree and the other one is in the University of Hong Kong with a computer science degree in 2013

Regarding my profile, I have 9 years working experience. Especially a total 5 years in data engineering area. After graduation, I had worked in the biggest Internet and IT company in Japan and Hong Kong. Currently I am working as a data director in Geely Group Geega Digital company.

I am familiar with the Hadoop ecosystem technologies such as Hive, Hbase, Spark, Flink, Elasticsearch, Neo4j,Clickhouse, Data lake Hudi and etc.

By using these tools, I have projects experience in building an enterprise level data warehouse, data lake in financial industry and automobile industry.

Additionally, I have cloud computing experience in AWS, Ali Cloud and Tencent Cloud. I also have experience to build a CDH system on-premises.

I also understand microservice Springboot, Spring Cloud and virtualization technologies such as Kubernetes and docker.

For the project management, I am using scaled agile Framework SAFe.

Apart from performing my daily duties, I made a point of turning my experience into innovations. In fact, I have authored 13 patents and one of them has been granted.

Also, in order to expand my skillsets, I studied and passed several technical certificate exams. For example, the AWS Certified Solutions Architect - Professional, DAMA Certified Data Management Professional, Project Management professional and a Sandford University machine learning certification that help me understand data science.

To sum up, I have solid technical skills in big data architecture, complex system design and development. Besides, I also have data governance practice.

HSBC is a world-famous financial institution and I really appreciate the opportunity of having this interview. I am confident that I can make strong contributions to the success of your team.

**Data Governance?**

Data Governance is about the roles, responsibilities and processes of ensuring accountability for and ownership of data assets.

The framework of DAMA includes Data architecture, data modelling and design, data storage and operations, data security and etc.

The benefit is to minimize risks, establish internal rules for data use, improve internal and external communication, increase the value of data, reduce costs

**Practice**: I tried a data governance in a Geely industry. Previously, they had a lot of equipment and systems. It was extremely hard for them to make full use of information; the decision chain is long and slow. I started from understanding their business and organization structure. And the second step is that I list all the benefits and ROI analysis to high level keyholders and asked them for an approve. The third step I did was to build a data platform to automatically collect data from some key data sources. The fourth step was that I build a BI screen for those high-level stakeholders to monitor key index of the industry in a real-time manner. Stakeholders realize the value of data governance then and encourage me to do more. Finally, I continuously get more and more data owners involved. Finally, I break the data island and achieve data-driven atmosphere in the industry.

**Why this company, this position?**

I would like to work for HBSC because of its fantastic reputation, mature management process and impressive track record of achievements. Everything you do is focused on the future success of the business.

I think my experience in data architecting is a great match with the position I applied. My previous experience will help me adapt to the new team and start to deliver strong performance quickly.

**Why do you leave your current company?**

I have been working with Geely Group for two years, as a big data director. Geely group is a great company and it gave me a lot of supports, and I gained so many skills. However, the amount of data in automobile industries is not big enough and the value of data is relatively small at current stage. I realized I was not challenged the way I used to be. Therefore, rather than let myself get too comfortable, I would like more challenging opportunity, which can make full use of my skills and take on more responsibilities.

**What is your short-term goal and long-term goal?**

For the short-term goal, I would like to integrate into the new team in a short time and complete a few projects to develop a trusting relationship with colleagues

In the next five years, I will steadily expand my skills and knowledge in big data by successfully completing projects and gaining new advanced skills. I would like to develop systems to help company accelerate business. Also, I want to become a source of knowledge and experience for junior members and help everybody in the team to succeed.

**What is your main achievement, strength?**

My strength is strong learning ability, self-motivation. I always keep learning the latest technologies and skills and use them in work to improve working efficiency.

I am action-oriented and methodical ( [məˈθɒdɪkl]). If I have a clear target, I always un-pact it into steps and achieve them one by one.

**What is your weakness?**

Sometimes, it is hard for me to say no to request and end up taking on more than I can handle. But I have realized that it is important to be realistic about my own time and prioritize my own tasks.

**Describe one of your project which shows your technical skills?**

Situation (time, where, what)

Challenges

Solution

Result

**(1)Geely Industry Internet Big data Platform project:**

Situation：1. In automobile industries, there are many kinds of data and abnormal data requires instant response. Daily incremental data is about 6 TB. Due to performance bottleneck, traditional Datawarehouse can not meet the requirement. Rich data forms lead to different data processing and data modeling methods. Therefore, a lot of big data technologies were utilized, the compatibility [kəmˌpætəˈbɪləti] of technologies is another challenge.

To solve the problems I mentioned, I utilized data lake Hudi for multiple data format processing and Clooudrea‘s CDH for big data modules management and maintenance. Thanks to this architect, 20 types of data could be collected successfully.

Another challenge in this project was the low data quality. TO solve this problem, firstly, I used Flink CDC for incremental data collection, then a ETL was used to do the data cleaning, feature engineering, data fusion and so on. Secondly, to ensure traceability of the data modelling, I utilized Open source framework Apache Atalas for the meta-data management and data lineage [ˈlɪniɪdʒ].

The last challenge of this project was the hardware resources management. In traditional data platform, resources is pre-assigned for computing and data storage. But peak and trough periods, the consumption is different, which means the waste. I used IaaS and PaaS services to achieve spark jobs dynamic scaling [ˈskeɪlɪŋ] for the large scale distributed system. By using these technology, 30% percent of resource cost was saved.

**(2) Geely Operation analysis system project**

Geely is lack of a system to do a real-time data analysis and decision chain is long, so its business decision speed is slow. The marketing and operation related data was stored in different systems. Therefore, I identified this business opportunity and designed a real-time Data lake for them.

By improving real time data analysis capability, I utilized Flink CDC connector. Change Data is Captured automatically from database bin log, event tracking system and message queue.

For the streaming ETL I used Kafka connect, Spark Streaming and Hudi. In this case, data was ingested by streaming pipelines and instant data analysis was achieved

The other challenge of this project was multiple sink joining and delayed data processing. I used flink watermark and broadcast.

The achievement of this project was 100,000 QPS at the peak, 500 Flink job processing at the same time and low latency.

**Adv of hudi:**





Community is more active.

**(3) Link & Co Customer data platform project**

Challenge is multiple-channel user mapping. In this project, the number of user was about 20 million that the calculation capacity is huge. I used Spark GraphX for the ID mapping

**Describe a project in bank industry?**

I worked at Chongqing Rural commercial Bank from 2018 to 2020. My main responsibility there were three tasks:

Firstly, I built an on-premise data warehouse for we chat bank credit business statement. I used star model for the modelling, dimensions including location, loan quality, bank branch and time.

Secondly, I integrated a few AI technologies such as ASR, TTS, face recognition，OCR and etc.. I integrated them to build bank check recognition system, voice and image dual recording system, headquarter visitor system and telephone outbound system.

Thirdly, as a technical leader, designed three bank-level loan products. I used micro services,

One of product ‘Ticket loan’ own a best practice prize in 2019 from 农信银资金清算中心.

I designed a architecture to accelerate notes discounted process. The whole process took 10 second to complete. For the traditional method, I took even days. The key point of my design was three reasons: 1. Re structured the note discount process, 2. Linked key systems and data in the bank, 3 utilized advanced technologies at that time. For instance, streaming ETL, federated learning which can union different bank, share information. Hence, it increased the credit approval time

**Architecture Consideration**

Designing a high-performance data warehouse architecture is a tough job and there are so many factors that need to be considered. Given below are some of the best practices.

* **Deciding the data model as easily as possible** – Ideally, the data model should be decided during the design phase itself. The first ETL job should be written only after finalizing this.
* At this day and age, it is better to use architectures that are based on **massively parallel processing**. Using a single instance-based data warehousing system will prove difficult to scale. Even if the use case currently does not need massive processing abilities, it makes sense to do this since you could end up stuck in a non-scalable system in the future.
* **If the use case includes a real-time component, it is better to use the industry-standard kppa architecture** where there is a separate real-time layer augmented by a batch layer.

**Handling Conflicts**

Open, honest, respectful communication

Be a good listener

Relate to the other person’s problems and see situation through their perspectives

Ready to make concessions

**Management experience?**

I have two parts of working: supervising junior engineers and technical architecture design.

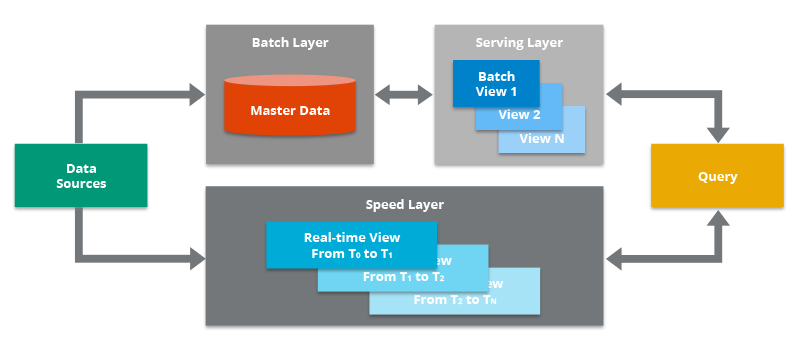
**For supervising employee**, I prefer to take a collaborative approach to working with and managing junior engineers. I would like to hear ideas, suggestions and concerns from them and help them by specific methods accordingly.

**For a data architect**, my task is a combination of creativity and practicality [ˌpræktɪˈkæləti]. Sometime I have a high-level vison and can think out-of-the-box. Sometimes, I participative in some specific function designs deeply and do coding for some core parts. It is important to be a bridge between business and technique. Therefore, it requires me to learn consciously. Besides, in big projects, excellent communication skills and persuasion [pəˈsweɪʒ(ə)n] ability are also crucial factors which affect the success of a project.

**Failure Experience?**

I was managing a project for one of our biggest clients and I was so eager to keep them happy. The client wished to end the project shorter than I estimated and refused to any negotiation. Even if I thought it was a risk, but I did not want to lose the client so I made my promise under pressure. Of course, I missed the deadline at the end and my client was disappointed. Looking back, I realized that I should pay more attention to the customer’s expectation control. I should unpack their business to serval small tasks and help them achieve many small success and finally add up to a big success.

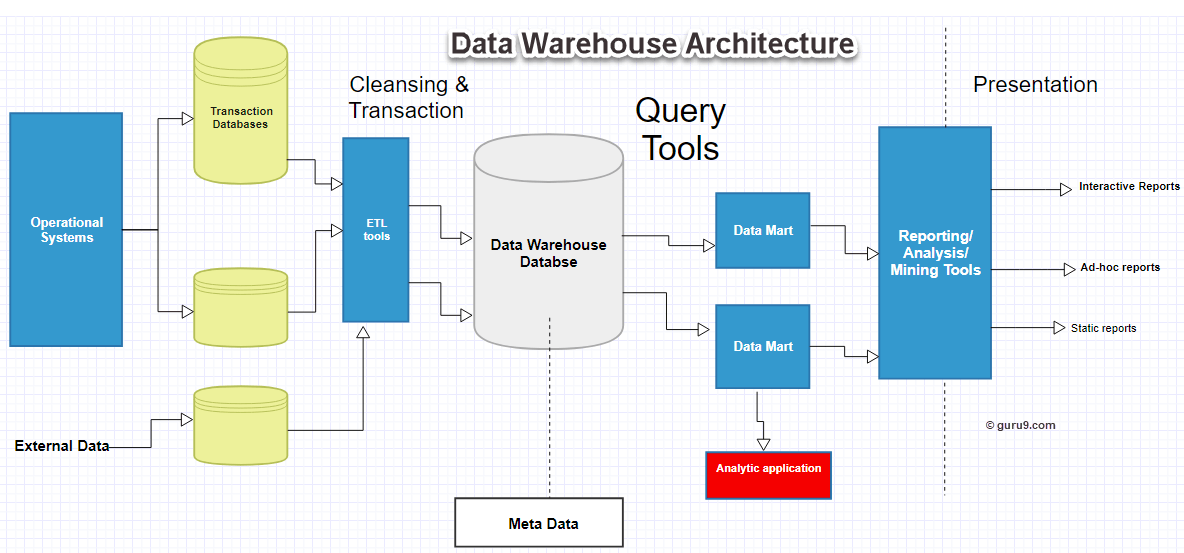
**Lambda Architecture?**



**Kappa Architecture**

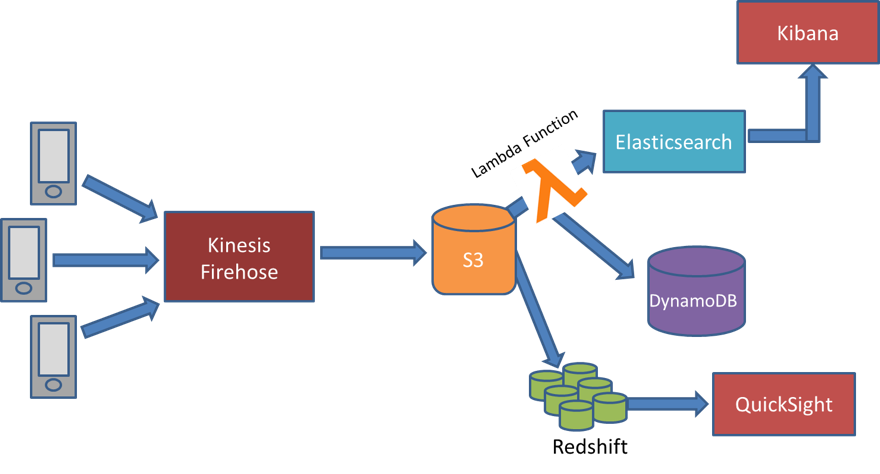
The main premise behind the Kappa Architecture is that you can perform both real-time and batch processing, especially for analytics, with a single technology stack.

**Data warehouse?**



**Real time data warehouse**

I built a RTDW using Cloudera to ensure a good customer experience and to keep maintenance costs under control.



Ingest the data, Aggregate the data, Analyze the data

**AWS experience?**

图示

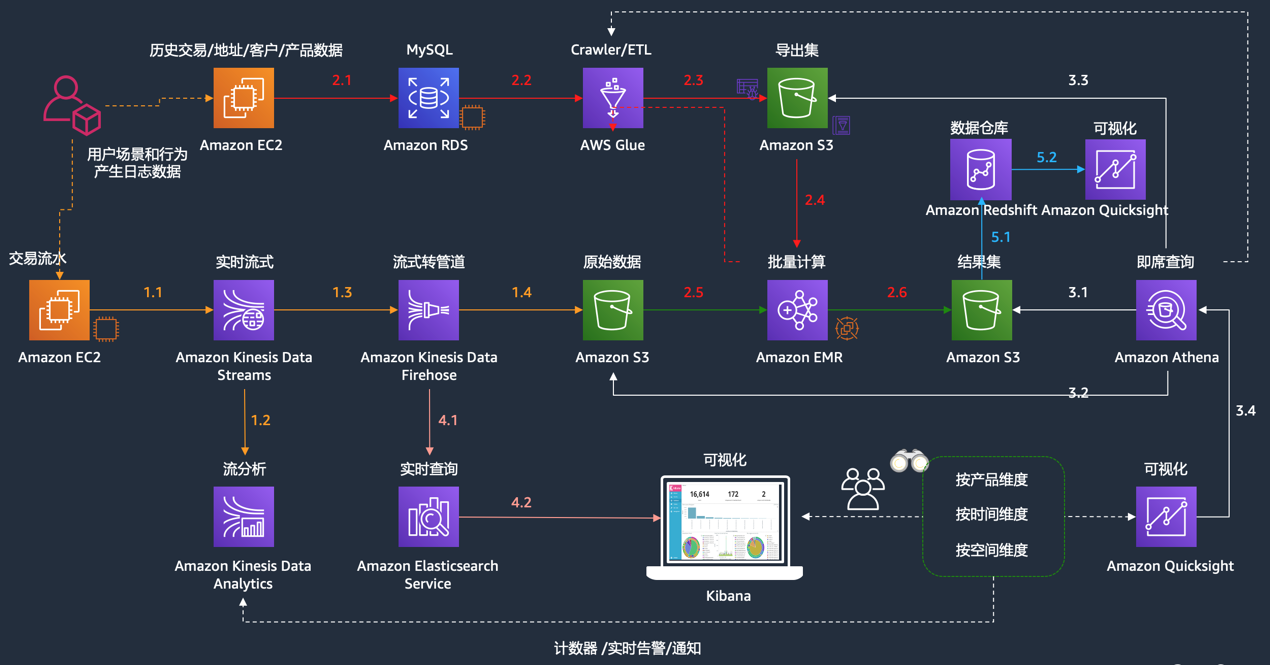
中度可信度描述已自动生成

[ɪnˈdʒest]

图示

中度可信度描述已自动生成

A project to do summary polar star customer transaction management and analysis.



Lab1：实时流数据处理，基于 Kinesis [kɪˈniːsɪs; kaɪˈniːsɪs] 产品家族实现

Lab2：批量数据处理，基于 EMR(Spark) 实现

Lab3：数据可视化，基于 Quicksight + Athena 实现

Lab4：数据实时检索，基于 Elasticsearch 实现

Lab5：数据仓库构建和数据可视化展现，基于 Redshift + Quicksight 实现

**hudi拥有2种存储优化。**

读优化（Copy On Write）：在每次commit后都将最新的数据compaction成列式存储（parquet）；

写优化（Merge On Read）：对增量数据使⽤⾏式存储（avro），后台定期将它compaction成列式存储。

**Question?**

What kind of candidate are you looking for? What does success look like for this position?

What is the biggest challenge for this position?