杜承涛

Mobile/WeChat: 15221516285 email: llangyyue@163.com

自我描述

- ✧ 阿里大模型证书，丰富的 AWS 和 Azure 云经验。
- ✧ 熟悉敏捷开发和项目管理，拥有 PMP 证书。擅长领导跨职能团队和多文化团队，具备出色的沟通能力和解决问题的能力，
- ✧ 能够与客户建立信任关系，准确理解他们的需求，并转化为实际的项目成果。
- ✧ 丰富的云计算、大数据、数据质量，数据血缘，数据合规、建模经验；能够从 0 到 1 进行系统的设计和开发。

个人作品：

AI 聊天工具：https://main.dnvvlhw0uwpc6.amplifyapp.com

任务管理工具：https://main.d1rfx5fdu5wom2.amplifyapp.com

图片处理 AI 工具：https://main.d2dysi9ch8qp9b.amplifyapp.com

技术：

- ✧ Scala、Python、Java& C#、React、UMI、Next Js、TypeScript、JavaScript、Tailwind、ShadCN
- ✧ 云计算平台（AWS\Azure\Vercel）、大数据组件（Hive、Spark、Kafka、Delta Lake）

工作经历

- ✧ 2021.03 – 至今 ：IBM 职位：Application Architect
- ✧ 2010.10 – 2021.02：美世（中国）有限公司 职位：DL：养老和退休基金系统的研发和设计
- ✧ 2008.07 – 2010.09：宽文软件 职位：工程师：仓储物流系统软件的开发和设计
- ✧ 2004.07 – 2008.06：北京盛安德科技有限公司 职位：工程师：Migration 部门的项目的开发

项目经验

IBM 公司项目：

2024.1 – 至今：TDC

项目描述：香港贸易发展局是香港法定机构，负责促进、协助和发展香港贸易 SME；主要通过举办国际展览会、会议以及商贸考察团协助企业；EOP 是参展商的参展平台，替换使用了 15 年的 legacy 系统；TDC 希望 EOP 能够支持本地和海外不同形式的 fair 和 conference，更加高效和具有弹性；

技术栈：

FE：Figma、Next.js、React、i18n、Zod、ShadCN、Tailwind

BFF：TypeScript、API Validator、Cache、Authentication

BE：Micro Service、Spring Boot、 Feign、ALB、JPA、Debezium、SonarQube、Mockito、JUnit

Cloud：S3、ECR、ECS、Fargate、Lambda、Api GateWay、SSM、SQS、MSK、Maria、Dynamo、Cloud Formation

- 主要职责

管理：根据整体项目的 target 和 scope，定义项目的 wave；依据不同的 wave，定义项目的 interation 的 target；与 Scrum Master 配合，进行项目的管理和沟通、完成每一次 iteration 中的项目进度和目标；
技术：FR 和 NFR 转化为技术解决方案，识别和解决技术问题，进行技术文档 HLD、LLD 的编写，技术的选型、功能模块和 API 的设计，服务的拆分、MR 的 Review；带领大连、北京和上海的前后端团队完成项目功能模块的开发。

2023.3 – 2024.12：大众（安徽）

- 项目描述：Smart Factory (IIoT)
- 主要职责：
  1. 运用云计算、大数据和安全技术，进行技术选型和架构设计，为客户定制对系统给出基于混合云的方案。
  2. 运用 IoT 技术，如 MQTT 协议等，构建设备与平台之间的稳定通信链路，确保数据的可靠传输和低延迟，保证设备数据的准确性和及时性。
  3. 根据项目需求，设计基于混合云的架构方案，合理规划资源分配。制定详细的项目计划，确保系统开发、测试和部署工作严格按照时间节点有序推进，保障项目顺利交付。
  4. 依据业务需求，开发定制化的 IoT 数据监控面板功能模块，如为不同客户角色设置权限管理、报表生成等。
  5. 与数据分析团队协作，对收集到的设备数据进行深度挖掘和分析，通过机器学习算法预测设备故障，提前进行维护提醒，降低设备停机时间。
  6. 为业务部门和客户提供平台使用培训，帮助他们熟悉 IoT 数据监控面板的各项功能，提升其自主管理设备的能力，同时收集用户反馈，持续优化平台功能和用户体验。
  7. 主动识别项目实施过程中的潜在风险，制定并执行有效的风险管理计划。定期开展风险评估，及时发现风险隐患，制定针对性的缓解策略，并持续监控风险状况，确保项目平稳运行。
  8. 负责编写和维护项目的技术文档，包括 prd，4a，hhd 和 lld 等。确保技术文档的完整性、准确性和及时更新，为项目团队提供全面的技术指导。严格完成技术文档的评审工作，保障文档质量，使其成为项目交付的重要支撑材料。

2021.3 – 2023.2：Nike GCDS（Great China Data Store）

- 项目描述：Nike 公司基于 AWS 的大数据平台，数据通过 DCS 系统和匿名系统摄取到数据湖中，结合数据质量、元数据管理、调度和 ETL 系统，进行数据存储、数据仓库建模以及数据的运算、衍生和分析，给 ML 和 BI 提供数据，给部门提供数据集市，给其它系统提供 API。
- 技术栈： Great Expectation, Spark、Kafka、Hadoop、Hive、Airflow、Delta Lake、Yarn、S3、Lambda,、CloudWatch、SQS、RDS、EC2、EMR。
- 主要职责
  - 设计：
  1. 梳理 Nike 的 Global 大数据平台框架，制定中国区的 Data Localization 路线图，对系统给出基于 AWS 云的方案，按照计划逐步进行系统的开发测试和部署。
  2. 设计和开发数据质量工具，包括设计和开发数据质量规则（强弱规则以及置信、SLA 等）、规则的验证和管理，数据质量报表、报警系统；并完成它与现有 ETL 系统的集成；解决集成过程中工具和现有 PIPELINE 融合出现的问题，做到集成后工程师对原有代码的无感知更新。

3. 元数据：设计和开发元数据收集和管理工具；收集技术元数据和业务元数据，提供管理和维护元数据的系统，元数据系统与其它系统的数据同步服务（Collibra 和 DQ）。
4. 根据 DSL 对公司数据合规的要求，根据 PAC 的规范和设计原则，整理和设计 DCU 系统，完成对数据的加密，脱敏，逻辑和物理删除，使数据系统符合数据合规的要求。使用 Delta Lake 完成数据在行级和列级的 CRUDM，数据版本控制和审计；
5. 与客户和其它 Vender 进行定期沟通，确保项目需求和目标得到统一和充分理解；对于执行过程中发现的问题，协调相关团队解决项目问题；在解决方案确定后，持续跟踪方案的落实和执行过程，确保所有行动和动作都按照既定时间表进行，并及时调整应对出现的新问题；
6. 识别和管理潜在风险并制定应对策略。通过定期风险评估和制定预防措施，确保了项目目标的实现不受意外干扰；建立了风险沟通机制，确保团队及时了解风险状态并采取相应行动，降低项目不确定性和潜在损失。

■ 技术：

9. 数仓：Nike 的 Consumer Behavior 数据通过 Sensor 先采集到 DSP 系统， 再通过 Spark Streaming 将 DSP 系统中的数据 Sync 到 DAS 系统，DAS 系统对数据进行冷热处理和归档，管理数据的全生命周期。在 Hive 中进行维度建模，RAW 层(采集存储用户的原始行为数据)，Cleansed 层(对数据进行去重，脱敏等数据清洗操作)，Curated 层(对数据按照分析维度进行聚合)，Semantic 层。 数仓各层之间的数据通过 HQL 等进行转换，通过 Airflow 进行作业的调度，把任务发送到 EMR 进行相关计算，并配置任务异常时发送 EMail 和 Slack 的 Alert，最终的计算结果 Sync 到 DAS 系统。数据质量通过 GX 进行 DAMA6 个维度的检查，并通过 Internal SLA(Airflow Built-In)和 External SLA(Lambda)，进行关键节点数据的检查，发送 Alert 给相关 BSA。数仓的数据血缘通过 Atlas，OpenLinage 和 Spline 结合 Marquez 进行 task level 和 data level 的采集和管理，可查看表和字段级别血缘，以及 task 和 dag 级别血缘，方便数据出现问题时，评估受影响的指标和 DAG，并进行问题追溯。
10. 数据质量：针对公司各业务域的数据质量问题，与 BSA，EDA 和 Data Localization 团队进行沟通调研，收集各业务域的数据质量方面的需求，根据需求进行数据质量工具的技术选型和设计，给出基于 AWS 的解决方案，完成 HLD 和 LLD，POC，MVP 和 Demo，制定开发和交付路线图，同 BSA 以及开发和测试团队协同工作，完成工具的开发，测试，部署、培训，交付。
11. 数据质量工具：
    使用 Next Js、Tailwind、ShadCN 开发。使用 AWS Lambda 调用 API。在 MongoDB 中保存和维护数据。将其部署到 AWS EC2。
12. 元数据收集工具：
    使用 React 和 Ant Design、Mock、Open API 设计和开发一个元数据收集和管理工具。使用 Spring Boot、Swagger 开发后端服务。使用 Redis 和 Okta 实现单点登录（SSO）。使用 Hive SQL 从 Hive、Delta Lake 收集元数据，并将它们保存到 MySQL 中。使用持续集成/持续部署（CICD）将其部署到 AWS EC2。

美世公司项目：

美世数仓：

系统架构：用户行为日志通过 Adobe 先采集到 Kafka，Kafka 起到解耦和流量削峰作用，再通过 Flume 将 Kafka 数据采集到 Hdfs 上。Myql 业务数据库数据通过 Sqoop 定时采集到 Hdfs。在 Hive 中进行维度建模，ODS 层(存储采集的原始用户行为数据和业务数据)，DWD 层(对数据进行有效性检验，去

重，脱敏等数据清洗操作，进行数仓维度建模，将业务总线矩阵中相关的事实表进行表 join 操作)，DIM 层(存储维度表数据)，DWS 层(对数据按照分析维度进行按天轻度聚合)，DWT(对数据进行重度聚合，统计近 N 天数据)，ADS 层(业务指标层，统计数据来源于 DWS，DWT，DWD，DIM)。 ADS 层(数据分析层)数据经过 Sqoop 导出到 Mysql。 数仓各层之间的数据转换脚本，通过 Azkaban 进行全流程定时调度，并整合配置任务异常时邮 件和电话报警。数据质量通过自定义 Shell 脚本实现数据完整性，一致性，指标波动等校验。日常临时指标统计，通过 Presto 即席查询工具实现。Hive 数仓的数据权限通过 Ranger 实现表，行，字段级别权限控制。数仓的元数据管理通过 Atlas 实现，可查看表和字段级别血缘依赖，方便任务失败时，评估受影响的指标及问题追溯。

- 主要职责:
  1. 搭建大数据集群，配置高可用，编写集群启停和数据分发脚本。搭建行为日志和业务数据采集通道，根据业务表的不同类型采用不同的同步策略，使用拉链表优化用户表的存储。
  2. 进行数据仓建模，同后台业务人员和产品经理沟通，梳理业务表关系和分析指标，根据数据库 ER 关系图，绘制业务总线矩阵，规划项目分层，在 Hive 中创建各层对应的表。
  3. 负责运营指标分析语句的编写，分析各项指标，如成交总额，转化率，活跃用户等，对 HQL 进行优化，处理开发过程中的数据倾斜和小文件等问题。
  4. 编写数据治理脚本，对数据进行过滤，脱敏处理，对数仓各层之间数据的完整性，一致性，时效性进行校验。
  5. 梳理重要的作业流程和数据血缘关系。
  6. 自定义 UDF、UDTF 等函数来解析公共字段和事件字段，使用窗口函数解决相关需求。

RTG（Retirement Technical Group）

- 项目描述：美世公司使用这个产品，为全球 300 多家客户提供退休金养老金和保险方案，以退休金管理为例，它可以支持大多数的欧美退休政策，例如经济条件假设，服务费，DBO 利息收入和支出，资产损益分析，福利模型，责任测量和资产上限责任等。包括的关键模块有，费用管理，财务披露，中期报告，财务责任披露灵敏计，资金和经济条件假设，未来财年数据，财务计算模版的审批和分析，批量财务和人事数据的管理。

  整个项目目前分为三个阶段；第一阶段：主要使用.NET 相关的技术实现上述项目中的功能模块的开发，并随着新客户的加入和各国政府福利政策的变化，更新原有功能或增加新的功能；第二阶段：随着数据量的不断增加和业务规则的更新，现有的引擎计算效率已经成为业务发展的瓶颈，公司决定把项目移植到云平台并使用大数据相关的技术（Spark、Hive、Hadoop、Kafka、Zookeeper），重构现有的业务逻辑，移植整个项目到大数据平台；第三阶段：建立数据仓库，通过数据驱动业务发展。
- 技术栈：Spark、Hive、Hadoop、Kafka、Zookeeper、Azkaban、Sqoop，S3、EC2、EMR、SQS、SNS、Lambda、RDS、CloudWatch
- 主要职责:
  1. 负责大数据平台设计、技术选型和研发，包括数据的接入、存储、计算、查询和分析，以及大数据相关标准制定；集群的规划、搭建与维护，保障大数据平台正常运行；
  2. 修改和调试 Spark 源代码，解决 Spark 使用过程中发现的 bug，对其进行二次开发；解决团队开发过程中的 Spark 的性能和数据倾斜等问题；
  3. 根据业务规则把不同的计算逻辑分别定义成不同的聚合函数，构建成不同的 Aggregator，实现自定义的算子，提供给团队使用；
  4. 自定义分区器，减少 Spark 使用过程中产生的 Shuffle；
  5. 从新规划、设计和优化系统中的原有模块重新设计和规划项目中的排序功能模块；
  6. 自定义数据结构实现项目中的分组 TopN 问题；

7. 设计，建设、维护和优化公司的数据仓库系统，进行数仓建模和开发。
8. 开发和设计系统的调度系统、元数据管理，数据质量管理，即时查询等；
9. 解决团队开发过程中的 Spark 的性能和数据倾斜等问题；
10. 分析并解决 Kafka 使用过程中数据重复消费和数据积压问题；
11. 解决 Hive 开发过程中的性能问题；
12. 参与和设计项目的架构，完成 POC 和 MVP 的设计实施和交付；
13. 与测试团队配合，研究适合公司大数据项目的自动化测试流程和制作相关工具；
14. 配合客户管理和分析数据，使得数据安全合规地在项目中被访问和使用。

上海宽文软件项目经历：

2008.07 – 2010.09：e-Logistics

- 项目描述：马士基公司仓储物流系统。
- 主要职责：完成项目的设计，同时参与开发和测试。
- 主要用到的技术：用来数据的采集和处理，例如各地仓储人员通过扫描抢在出库入库时扫描货物信息，并通过无线或者有线的方式传输到服务器上,部署在各个仓库服务器上的服务会处理数据，并汇集到数据中心。在此过程中，项目使用了多种查询排序算法来处理上传上来的数据,并且使用各种事务处理技术防止并发。采用了微软自带的多语言技术，并结合项目特性,给客户提供了特定的多语言解决方案。

北京盛安德公司项目经历：

2004.03 – 2008.06：AppSetter + DCO + TRO

- 项目描述：该 AppSetter 是一个在线日程管理工具，允许业主在线安排一个与设计中心随时随地的会议，当安排会议的时候，用户可以发送电子邮件通知每个参与者。DCOps 是一个在线的计划和沟通工具；用户可以选择来自产品数据库的产品来确定建造或装修的时候使用什么样的产品。一旦用户选择的产品确定下来，就为用户打印合同，同时生成采购订单；当一个采购订单生成后，系统生成电子订单分发给对应的供应商，同时系统会生成一个建造和安装计划，每一个安装和建筑公司收到这个计划，就会按照计划中定义的时间点来进行它们的工作。该工具还允许建造商和生产商上传产品，设计师设计图纸，以便用户直观的选购产品并查看设计师的设计。所有合同和工作订单上的数据（成本，劳动量）有系统自动计算。这个系统有助于使建造商，设计师，生产商相互配合，使得整个房屋的建造或装修过程非常高效。TradeOps 是一个调度和跟踪工具；建筑物的最后订单生成后，这个工具可以安排调度工作人员，跟踪产品生产和加工情况，跟踪保养和维修工作，保证后期工作的效率。建造维修工作，采购订单和电子客户服务代表等是通过电子邮件确认收到各自的工作并确认完成情况；业主能够随时随地查看自己房屋的建造维修情况。
- 主要职责：同客户沟通，进行前期需求分析，协同架构师完成项目的设计，同时参与开发和测试。

教育背景

- 2000.09 – 2004.07 东北林业大学 本科 学士学位 专业：计算机科学与技术

Self-Description

- Familiar with new retail industry and automotive industry.
- Holder of Alibaba Large Model Certificate, with extensive experience in AWS and Azure clouds.
- Familiar with agile development and project management, holder of PMP certificate. Skilled in leading cross-functional and multi-cultural teams, with excellent communication and problem-solving abilities.
- Capable of building a relationship of trust with customers, accurately understanding their needs and translating them into concrete project outcomes.
- Rich experience in cloud computing, big data, data quality, data lineage, data compliance, and modeling; able to design and develop systems from scratch.

Personal Projects

AI chat tool: https://main.dnvvlhw0uwpc6.amplifyapp.com

Task management tool: https://main.d1rfx5fdu5wom2.amplifyapp.com

Image processing AI tool: https://main.d2dysi9ch8qp9b.amplifyapp.com

Technologies

Scala, Python, Java & C#, React, Next Js, TypeScript, JavaScript, Tailwind, ShadCN

Cloud platforms :(AWS, Azure, Vercel), big data components (Hive, Spark, Kafka, Delta Lake)

Work Experience

- IBM (Mar 2021 – Present): Application Architect
- Mercer (China) Co., Ltd. (Oct 2010 – Feb 2021): Leader of the Development
- Kanwen Software (Jul 2008 – Sep 2010): Engineer
- Beijing Shengande Technology Co., Ltd. (Jul 2004 – Jun 2008): Engineer

Project Experience

IBM Projects

Jan 2024 – Present: TDC

Project Description

The Hong Kong Trade Development Council is a statutory body in Hong Kong. It promotes, supports, and develops Hong Kong's trade SMEs. It mainly assists businesses by organising international exhibitions, conferences, and trade missions. The EOP is an exhibition platform for exhibitors, replacing a legacy system

that had been in use for 15 years. TDC wants EOP to support various local and overseas fairs and conferences, making it more efficient and flexible.

Technical Stack

FE: Figma, Next.js, React, i18n, Zod, ShadCN, Tailwind
BFF: TypeScript, API Validator, Cache, Authentication
BE: Micro Service, Spring Boot, Feign, ALB, JPA, Debezium, SonarQube, Mockito, JUnit
Cloud: S3, ECR, ECS, Fargate, Lambda, Api Gateway, SSM, SQS, MSK, Maria, Dynamo, Cloud Formation

Key Responsibilities

Management

Define project waves based on overall target and scope. Define iteration targets for different waves. Work with the Scrum Master to manage the project, communicate, and complete goals in each iteration.

Technology

Translate FR and NFR into technical solutions. Identify and resolve technical issues. Write technical documents (HLD, LLD). Select technology, design functional modules and APIs, decompose services, and review MRs. Lead front-end and back - end teams in Dalian, Beijing, and Shanghai to develop project functional modules.

March 2023 – December 2024: Volkswagen (Anhui)

Project Description: Smart Factory (IIoT)

Key Responsibilities

- Utilize cloud computing, big data, security and technologies to conduct technology selection and architecture design, and customize hybrid cloud-based system solutions for clients.
- Use IoT technologies, such as the MQTT protocol, to establish stable communication links between devices and platforms, ensuring reliable data transmission, low latency, and the accuracy and timeliness of device data.
- Design architecture schemes based on hybrid clouds according to project requirements, plan resource allocation effectively, formulate detailed project plans, and ensure that system development, testing, and deployment proceed on schedule to guarantee smooth project delivery.
- Develop customized IoT data monitoring panel functional modules based on business requirements, such as setting up permission management and report generation for different customer roles.
- Collaborate with data analysis teams to deeply mine and analyze collected device data. Use machine learning algorithms to predict equipment failures, provide advance maintenance alerts, and reduce equipment downtime.
- Conduct platform usage training for business departments and customers to familiarize them with the functions of IoT data monitoring panels. Improve their ability to independently manage equipment, collect user feedback, and continuously optimize platform functions and user experience.

- Actively identify potential risks during project implementation, develop and execute effective risk management plans. Conduct regular risk assessments, detect risks in a timely manner, develop targeted mitigation strategies, and continuously monitor risk conditions to ensure stable project operations.
- Responsible for writing and maintaining project technical documents, including PRD, 4A, HHD, and LLD, etc. Ensure the completeness, accuracy, and timely updates of technical documents, provide comprehensive technical guidance for project teams, strictly complete technical document reviews, ensure document quality, and make them important supporting materials for project delivery.

Nike GCDS (Great China Data Store) Project (Mar 2021 – Dec 2023)

Project Description

Nike's big data platform based on AWS. Data is ingested into a data lake via the DCS system and an anonymous system. The platform combines data quality, metadata management, scheduling, and ETL systems to perform data storage, data warehouse modeling, and data computation, derivation, and analysis. It provides data for ML and BI, offers a data mart to departments, and provides APIs to other systems.

Technical Stack

Great Expectation, Spring Boot, Java, Airflow, Yarn, S3, Lambda, CloudWatch, SQS, RDS, EC2, EMR.

Key Responsibilities

Management

- Clarified the framework of Nike's Global Big Data Platform, established a Data Localization roadmap for the China region, and provided an AWS - based system solution. Conducted system development, testing, and deployment as per the plan.
- Communicated with clients and other vendors to ensure a unified understanding of project requirements and goals. Coordinated with relevant teams to address issues arising during project execution. Tracked the implementation of solutions to ensure all actions were carried out according to the timetable and adjusted for new problems.
- Managed project requirements using agile development methods, ensuring accuracy and relevance. Established a change management process for requirements to handle dynamic changes and minimize impacts on project progress and budget. Ensured project delivery on time and within budget.
- Identified and managed potential risks developed response strategies. Conducted regular risk assessments and established preventive measures to avoid project goal disruptions. Set up a risk communication mechanism to keep the team informed of risk status and reduce project uncertainty and potential losses.

Technical

- Data Warehouse: Consumer behavior data was collected via sensors into the DSP system and synchronized to the DAS system using Spark Streaming. The DAS system managed the entire data lifecycle. Dimensional modeling was performed in Hive across multiple layers, with data transformations between layers using HQL. Airflow was used for job scheduling and task execution on EMR. Data quality

was monitored using GX across six DAMA dimensions, with alerts sent via Internal SLA (Airflow Built-In) and External SLA (Lambda). Data lineage was managed using Atlas, Open Linage, Spline, and Marquez.

- Data Quality: Collaborated with cross-functional teams to gather data quality requirements and designed technical solutions based on AWS. Completed HLD and LLD, POC, MVP, and Demo. Developed a roadmap for development and delivery, and worked with teams to complete tool development, testing, deployment, training, and delivery.
- Data Quality Tool: Developed using Next Js, Tailwind, ShadCN. Used AWS Lambda for API calls, stored and maintained data in MongoDB, and deployed on AWS EC2.
- 12. Metadata Collection Tool: Designed and developed using React, Ant Design, Mock, and Open API. Developed backend services with Spring Boot and Swagger. Implemented SSO using Redis and Okta. Collected metadata from Hive and Delta Lake using Hive SQL and stored it in MySQL. Deployed using CI/CD on AWS EC2.

Mercer Projects:

Mercer Data Warehouse

Architecture: User behavior logs are first collected by Adobe into Kafka, which acts as a decoupling layer and traffic peak shaving. Then, Flume collects Kafka data into Hdfs. Myql business database data is periodically collected into Hdfs via Sqoop. Dimensional modeling is performed in Hive, with the following layers: ODS (stores raw user behavior and business data), DWD (data validation, deduplication, masking, etc., dimensional modeling, table join operations), DIM (stores dimension table data), DWS (light - aggregation by day), DWT (heavy - aggregation, statistics for the past N days), ADS (business metrics layer, data from DWS, DWT, DWD, DIM). ADS layer data is exported to Myql via Sqoop. Data transformation scripts between layers are scheduled by Azkaban, with email and phone alerts for task exceptions. Data quality is checked via custom Shell scripts for integrity, consistency, and metric fluctuation. Daily ad - hoc queries are done via Presto. Hive data permissions are controlled by Ranger at table, row, and column levels. Metadata management is done via Atlas, showing lineage dependencies for tables and columns, aiding in assessing impacts and tracing issues.

Key Responsibilities

1. Set up big data clusters, configure high availability, write scripts for cluster start - stop and data distribution. Build collection channels for behavioral logs and business data, using different sync strategies based on table types, optimizing user table storage with snap - on tables.
2. Perform data warehouse modeling. Communicate with back - end business staff and product managers, clarify business table relationships and analysis metrics. Draw business bus matrix based on database ER diagram, plan project layers, create corresponding tables in Hive.
3. Write operation metric analysis statements, optimize HQL, handle data skew and small file issues in development.
4. Write data governance scripts for filtering and masking, verify data integrity, consistency, and timeliness across layers.
5. Document key workflows and data lineage.
6. Develop custom UDFs and UDTFs to parse common and event fields, use window functions to meet requirements.

RTG (Retirement Technical Group)

Project Description: Mercer uses this product to provide retirement, pension, and insurance plans for over 300 global clients. It supports most US and European retirement policies like economic assumptions, service fees, DBO interest, asset analysis, benefit models, liability measurement, and asset - liability management. Key modules include cost management, financial disclosure, interim reporting, financial liability sensitivity, funding and economic assumptions, future - year data, financial template approval, and batch financial/personnel data management.

The project is divided into three phases. Phase 1: Develop functional modules using .NET technologies, update features for new clients and policy changes. Phase 2: Migrate to the cloud and use big data technologies (Spark, Hive, Hadoop, Kafka, Zookeeper) to restructure business logic and move to a big data platform due to engine computation bottlenecks from increasing data and updated rules. Phase 3: Build a data warehouse to drive business with data.

Technical Stack

Spark, Hive, Hadoop, Kafka, Zookeeper, Azkaban, Sqoop, S3, EC2, EMR, SQS, SNS, Lambda, RDS, CloudWatch

Key Responsibilities

1. Design big data platform, select technology, develop data processing capabilities, set standards. Plan, build, and maintain clusters to ensure platform operation.
2. Modify and debug Spark source code, fix bugs, enhance performance, solve data skew issues.
3. Define custom aggregators based on business rules, create operators for team use.
4. Develop custom partitioners to reduce Spark shuffling.
5. Redesign and optimize system modules and sorting functions.
6. Design data structures for grouped TopN problems.
7. Design, build, maintain, and optimize data warehouse systems, perform modeling and development.
8. Develop scheduling, metadata management, data quality management, and instant query systems.
9. Solve Spark performance and data skew issues.
10. Analyze and resolve Kafka data repetition and backlog problems.
11. Address Hive performance issues.
12. Participate in project architecture design, deliver POC and MVP.
13. Collaborate with testing teams on automated testing for big data projects.
14. Assist clients in data management and analysis, ensuring secure and compliant data access and use.

Educational Background

Northeast Forestry University (211 University)
Bachelor
Computer Science and Technology