魏晓鹏

(+86) 177-2460-1381 wxp_sampson@163.com https://simonwei97.github.io linkedin.com/in/xiaopeng

教育背景

南方科技大学 2015.9 - 2019.6

- GPA: 3.41/4.00。**通信工程**，工学学士。

工作经历

商汤科技 | 软件开发工程师 2020.7 - 至今

- 负责基于知识库的 LLM ChatBot QA 系统中**检索服务**，完成从 0 到 1 的技术方案设计，最终实现工程落地。
- 主导 SenseFoundry 产品的视图设备管理服务、视图接入服务的开发工作，包括技术方案、技术文档撰写，代码开发，组件 Helm Chart 修改。完成了 7 个版本的开发工作，并且在职期间申请 2 个行业相关专利。

商汤科技 | 测试开发工程师 2019.7 - 2020.6

- 负责 SenseFoundry 产品设备管理服务、图片接入、视频接入服务的版本测试工作，包括测试方案制定、自动化脚本开发及测试报告输出。
- 使用 Python 开发自动化测试脚本，完成产品服务的功能、业务流、准确性及性能测试，共计完成了 3 个版本的测试工作。

项目经历

SenseNova Mini - Semantic Retrieve | *Python, FastAPI, Minio, Milvus, TiDB* 2023.8 - 至今

- LLM 知识库检索服务 (Semantic Retrieve)，接收用户的问题，Embedding 用户问题后，在 Milvus（文本向量库）中匹配出与问句向量最相似的 top_k 个，再对 top_k 结果进行上下文召回，最后将检索召回的文本作为上下文（context）和问题构造合适的 Prompt，发送给 LLM 生成回答。其中文本向量是由解析服务解析不同文档，之后切分为 chunk 文本，Embedding 后的结果。
- 服务基于 *Uvicorn* + *FastAPI* Web 框架进行开发。支持使用 Helm Chart 部署在 k8s 集群。
- 针对向量检索结果，设计了基于不同 chunk 文本类型的上下文召回策略，以获取更多相关文本给到 LLM，提升回答的准确性。整个系统解析和检索支持诸多文件类型，如 PDF/Doc/Docx, Q&A 问答对。
- 完成 CPU 版本 Milvus 在 x86 和 ARM 平台单机部署 (standalone) 形态的性能摸底测试。

SenseFoundry -IIS | *Go, gRPC, Zookeeper, OSS, Redis, Kafka, Prometheus+Grafana* 2020.7 - 2023.7

- 视图接入服务（IIS），接收 SDK 相机及其他视图协议的消息，将不同数据结构的对象消息过滤、清洗为指定内部结构，最后写入到 Kafka 消息队列，由视图解析服务消费后进行特征等细节分析。
- 服务基于 *gRPC* 开发，使用了 Zookeeper, OSS, Redis, Kafka/MQTT，使用 Helm Chart 部署在 k8s 集群。
- 服务采用自研 OSS 存储部分对象大图，然后将消息写入 Kafka（带小图或大图)。后续因业务场景越来越复杂，大消息传输时 Kafka 出现性能瓶颈，将消息中较大的图片的写入 Redis 缓存，解析服务消费后拉取后从 Redis 获取缓存图片，以低 Kafka 存储压力，改造工作将服务的接入处理能力提升了 30%。
- 实现链路数据治理，在输入到输出各环节 metrics 埋点，使用 Prometheus+Grafana 实现可观测性监控。
- 开发 CLI 工具，支持服务的性能测试，Mock 上下游服务及数据，运维问题定位、异常检测等。

SenseFoundry - DMS | *Go, gRPC, TiDB, Prometheus+Grafana* 2020.7 - 2023.7

- 视图设备管理服务（DMS)，将相机、NVR、IoT 等设备接入系统后统一进行管理，以 OpenAPI 形式提供接口对外。主要功能包括设备管理，平台管理，设备任务管理。支持多租户（数据库分表)。
- 服务基于 *gRPC* 开发，使用 TiDB（前期 MariaDB）进行数据持久化存储，采用 Helm Chart 部署在 k8s 集群。

专业技能

- 熟悉 Go, Python 开发，熟悉 gRPC 框架，GC 垃圾回收策略，GMP 调度，熟练使用 pprof 分析定位问题。
- 熟悉 Kafka 的高性能、高可用，负载均衡策略，熟悉 Kafka partitions/replicas 分配策略。
- 了解 Redis 内存淘汰策略，Redis 分布式锁，Redis 持久化。熟悉 Milvus 向量数据库的使用。
- 熟悉 Docker 及 Kubernetes 生态，有使用 Helm Chart 部署产品服务经验；了解 k8s pod 调度策略，熟悉 k8s 服务发现；有使用 Prometheus+Grafana 完成服务可观测性的项目经验。
- 大学英语四/六级（CET-4/6)，熟悉英文读写。具备英语日常交流的能力。