Q. RAG 구성을 위한 하드웨어 스펙이나 아키텍처 구성은 어떻게 되는가?

A. 우선 고객이 적제하고자 하는 데이터의 현황 및 용량을 확인하여 거기에 맞는 하드웨어 구성이 필요합니다. RAG 구성의 경우 온프레미스에 LLM을 구축하는 만큼의 하드웨어 스펙이 필요하진 않으며 1대 서버에 1.4TB 저장하는기준으로 DataNode를 구성하였을 때 1대 Data Node는 16 Core 64G RAM SSD 2TB 가 필요합니다.

스케일아웃 구성으로 목표 용량에 따라 Data Node를 구성하며 관리형 Node와 전처리 수집 Node는 분리하는것이 좋으나 사업 규모와 구성에 따라 통합 구성 하는 경우도 있습니다.

Q. RAG 기반 챗봇 개발 기간은 대략 얼마나 되는가?

A. 보통은 4개월에서 5개월 정도 소요 됩니다.하지만 고객의 데이터 준비 상황과 납기 일정에 따라 빠르면 3개월에도 구축 가능하나 인력이 많이 소요 되며 적극적 고객 지원 및 스콥 관리가 필요한 부분이 있습니다.

Q. RAG 말고 DB 데이터를 조회하여 처리 지원 가능한지? (Text to SQL 관련)

A. 멀티에이전트 구성 시 RDB에 JDBC 방식으로 연계하여 쿼리를 통하여 데이터를 가지고 올 수 있도록 SQL Query Agent를 만들 수 있습니다, 해당 에이전트를 와 LLM 연계를를 통해 Text-to-SQL 방식의 응답도 지원 가능합니다.