**[SSAFY] 8기 멘티 활동 일지**

**작성일자: 2023년 03월 20일**

|  |  |  |  |
| --- | --- | --- | --- |
| **도메인** | **빅데이터 분산** | | |
| **멘토** | **홍승환** | **팀 코드** | **C206** |
| **진행 일시** | 2023년 03월 20일 13:40 ~ 14:10 | | |
| **진행 내용** | 머신러닝 연구와 분산 데이터 처리에 대한 연구를 수행 중이신 스캐터랩의 홍승환 멘토님과 프로젝트 진행에 관련된 두번째 면담을 진행할 수 있었습니다. 저희가 드렸던 사전 질문들에 대해, 다음의 답변들을 얻을 수 있었습니다.  **Q1. HDFS에 저장된 분석 결과 데이터를 외부 RDBMS에 변환하여 제공하는 것이 Sqoop의 역할로 이해했습니다. 그렇다면, Spark에서 분석된 결과를 HDFS에 저장하지 않고 Spring으로 직접적으로 전송해서 RDBMS로 바로 Insert하면 될 것 같은데, 그렇다면 Sqoop은 필요하지 않는건지 궁금합니다.**  A1. 맞다. Sqoop은 분산 처리된 데이터를 고성능 소스로 옮겨주는 역할을 수행한다. 따라서, Spark로 처리한 결과를 바로 RDBMS로 전송할 경우, Sqoop은 필요하지 않게 된다. Sqoop은 더 이상 개발되지 않는 프로젝트이며, 현재 다양한 대체재들이 개발되고 있다.  **Q2. 현재 저희가 기획한 ERD에서는 대용량의 데이터를 효율적으로 처리하기 위해 크롤링한 기사 데이터(기사 제목, 내용, 기자, 언론사)를 모두 HDFS에 저장하고 있습니다. 이때, 홈페이지에 뉴스를 뿌려주기 위한 실시간 조회에 HDFS를 사용해야할지, RDBMS를 사용해야할지 궁금합니다. 이전 기수들이 수행했던 프로젝트를 참고했을 때, 2.8억 개의 데이터를 조회하는데 Hadoop을 적용하지 않았을 때 25초가 소요되고, 적용했을 때 0.6초가 소요되었다는 정보를 확인할 수 있었습니다. 다만, 저희가 Hortonworks에서 Hive를 통해 소규모의 데이터를 단순 조회했을 때는 10초가 넘는 시간이 소요되었습니다. 이러한 차이가 데이터 노드를 병렬 지정하지 않아서 그런 건지, 아니면 HDFS가 실시간 응답에는 적절하지 않기 때문에 정형화된 기사 데이터를 RDBMS에 저장해야하는 건지 궁금합니다.**  A2. 이전 기수의 결과는 나올 수가 없다. 이미 한 번 처리된 분석 결과를 캐싱해서 쓴 건 아닐까라는 강한 의구심이 든다. Hive에 단순 질의를 수행했을 때 10초가 넘는 결과가 나오는 것이 지극히 정상적이며, 당연하게도 RDBMS에 기사 데이터를 저장하고 사용해야 실시간 응답성을 보장할 수 있다. Columnar Database와 OLTP(Online Transaction Processing) vs OLAP(Online Analytical Processing)에 대해서 좀 더 알아보면 좋을 것 같다.  **Q3. 저희는 기사의 유의미한 통계 정보를 표현하기 위해 워드 클라우드, 트리맵, 원형 차트 등을 고려하고 있습니다. 해당 도구들 외에, 현업에서 사용하시는 시각화하기 좋은 도구가 있는지 질문드리고 싶습니다.**  A3. 대화형 인공지능 직군에서 데이터를 분석하기 위해서 별도로 시각화를 하지는 않고, Pandas 라이브러리를 이용하여 분석 결과를 바로 이용하는 편이다. 홈페이지 시각화를 위해서, D3.js를 이용한다면 데이터를 표현할 수 있는 여러 방법들을 알 수 있을 것이다. 파이썬의 경우, 장표를 시각화 하기 위해 Seaborn을 이용하는 것도 좋은 방법이다.  [참고 사이트]  <https://observablehq.com/@d3/gallery>  **Q4. Spark SQL로 Hive를 대체할 수 있는지 궁금합니다.**  A4. Hive는 데이터 저장 기능이 존재하는 데이터 웨어 하우스이고, Spark SQL은 데이터 엔진 위에 질의할 수 있는 라이브러리로 명확하게 구분된다. 처리한 데이터를 굳이 저장할 필요가 없는 경우, Spark SQL만 사용해도 무관하다.  **Q5. 저희 팀이 기획한 데이터 흐름이 적절하게 구성되었는지 질문드리고 싶습니다. (23p)**  A5. 적절하게 구성된 것 같다. 다만, 여기서 약간 애매한 것은 실시간 크롤링을 예시로 들었을 경우에 한정된다는 것이다. 배치 처리가 사용되는 경우에는 다르게 구성해야한다.  **Q6. 현업에서 빅데이터와 관련된 개발을 진행하시면서 자주 발생했던 에러나 그에 대한 대응책이 궁금합니다.**  A6. 업무를 수행하다 보면 TB단위로 처리하는 경우가 일반적이다. Group By 질의를 했을 때, 처리해야할 데이터의 규모가 너무 방대한 경우 RAM 부족으로 인해 OOM(Out Of Memory) 에러가 발생하게 된다. 해당 에러를 피하기 위해서, 첫 번째 방법은 서버 증축 및 성능 고도화를 채택할 수 있고 두 번째 방법은 기존 질의보다 좀 더 세부적으로 쪼개어 질의 하는 것이다. 예를 들어, 지역(광역시)으로 대상으로 한 질의에서 OOM이 발생한 경우 지역(시,군,구)을 대상으로 질의 범위를 좁힌다면 에러를 회피할 수 있게 된다.  **Q7. (추가질문) 저희 팀이 기존에 기획했던 데이터 흐름은 실시간 크롤링 처리에 적합하다고 말씀해주셨는데요, 그렇다면 크롤링 데이터를 배치 처리했을 때는 아키텍처를 어떻게 구성하는게 적절할지 질문 드립니다.**  A7. 배치 처리를 할 때는 크롤러가 크롤링한 데이터를 직접 HDFS에 적재한 다음, Hadoop에서 해당 데이터들을 처리하면 된다. 즉, Kafka의 필요성이 떨어지게 된다.  **Q8. (추가질문) 현업에서 프로젝트 관리를 할 때 JIRA외에 사용하시는 협업 도구가 있는지 궁금합니다.**  A8. 현업에서는 JIRA보다는 Notion과 Slack을 연동하여 사용하는 편이다. SW 자동 배포 혹은 장애 처리, 실시간 응답은 Slack Bot을 통해 해결하는 챗옵스(ChatOps) 방법을 도입하여 사용하고 있다. | | |