[SSAFY] 8기 멘티 활동 일지

작성일자: 2023년 3월 9일

<table>
<tr><td>도메인</td><td colspan="3">빅데이터 분산</td></tr>
<tr><td>멘토</td><td>서일환</td><td>팀 코드</td><td>B308</td></tr>
<tr><td>진행 일시</td><td colspan="3">2023년 3월 9일</td></tr>
<tr><td>진행 내용</td><td colspan="3">도메인 질문

1. 빅데이터 관리는 데이터 사이언티스트에 국한된 직무인가요?

- 데이터 엔지니어라는 직무가 생긴지 오래 안됨
- 빅데이터라는 분야가 집대성되기 시작한 것은 2014년 정도부터, 에코시스템들이 나오기 시작
- 멘토님은 웹 개발자일 때 데이터 엔지니어의 활동을 해왔음
- 데이터를 미리 연산해서 대시보드를 만들었음
- 데이터 분석 체계를 만드는 것이 중요함

2. 하둡에서 스쿱을 통해 RDB에 적재해서 간접적으로 사용하는 것으로 알고 있습니다. HDFS에 직접 접근하는 방법도 있을까요?

- 접근방법은 있지만 꼭 실무에서 꼭 필요는 없다. 그래도 경험상 이용해보면 좋다. 어떤 노드에 어떤 식으로 데이터가 들어가는지 실습해보면 좋음.

3. 하이브쿼리가 맵리듀스 알고리즘을 대신 사용해주는 것으로 알고 있는데, 맵리듀스 알고리즘 현 방식에 대한 학습이 필요하다고 생각하시나요?

- 학습에는 의미가 있음. 신입 데이터 직무에는 하둡 사용경험보다는 hdfs를 이해하고 알고 있는지, 알고리즘 작동 원리를 아는지 등을 더 중요하게 생각함.

4. 분산처리 알고리즘은 모두 정형화되어 있나요?
- 어느 정도 정형화, 응용 프로그램 개발자는 분산 처</td></tr>
</table>

리 개념을 알고 있다면 개발된 툴 사용 방법을 익히는 것도 좋음.

현업 관련 질문

1. 하둡은 실시간 분산처리가 어렵다고 알고 있습니다. 프로젝트에 적용하기에는 스파크를 이용하는게 효율적일까요? + 현업에서 하둡을 자주 사용하나요?

   - 하둡은 적재에 최적화, 값싼 기술, 하둡 기본 생태계만 써도 크게 문제 없음
   - 스파크는 실시간과 배치 처리할 때 유용, 스파크는 실시간에 특화
   - 데이터 인풋이 계속 있을 때, 얼마나 짧은 단위를 자르고 잘린 데이터를 빠르게 처리하는데 스파크 이용
   -

2. 실무에서 Cloudera Manager를 사용하나요?
   - 팀 by 팀, 하둡 인프라를 관리하는 곳에서는 많이 사용함

3. 현업에서 데이터 크롤링을 많이 하시나요?
   - 원티드(멘토님 현 직장)에서는 전세계의 채용공고를 크롤링하고 분산처리하긴 함
   - 실시간 데이터는 크기가 그리 크지 않은 경우가 대부분이라서 데이터 처리 파이프라인은 단순하게 만들어짐

4. 현업에서는 일반적으로 어느 정도의 크기여야 빅데이터 분산처리 시에 의미가 있을까요?

   - RDB에서 처리 가능한지가 중요함. 초당 데이터베이스에 들어오는 데이터량에 따라 분산 고민을 해야함.
   - 연산(cpu코어)시간 문제보다는 스토리지와 메모리 사이즈문제, 수 초 안에 끝나면 분산 필요없음, 테라바이트 단위, 하나 머신으로 불가능하면 스파크 같은 솔루션
   - 디스크 스캔으로 연산하는 시간이 충분하다면 분산 시스템은 굳이 필요하지 않음

	프로젝트 관련 질문 1. 데이터 마이닝을 실시간으로 보여주면 매번 데이터를 보내줘야 하는지? - 이전 100개의 데이터를 연산한 것을 보여주고 새로 추가되는 내용이 있다면 다시 연산 후 새로운 것을 보여줌. 2. 프론트에서 데이터를 보관하고 있으면서 request를 줄여야 하는지 아니면 매번 request를 해야하는지? - 규모가 작으면 굳이 고민할 필요는 없음, 다만 나중에 빠르게 처리하기 위한 전환 준비는 필요하다. 자유질문 1. IOT 관련하여 하둡을 얹어서 사용하는 경우가 있는지? - 발행되는 데이터를 하둡에서 처리하는 경우는 있을 수 있음. 실용적 차원에서는 없을 듯 2. 현업에서 카프카 이용하는지? - 카프카 이용함. 데이터가 저장되는 속도보다 빠르게 발행될 때 사용하면 됨.