**빅데이터프로그래밍**

**최종보고서**



60161592 김범석

60171759 장동혁

60191642 고성인

**목차**

1. 주제 및 목표 소개

2. 시스템 아키텍쳐

3. 데이터 분석 과정

4. 결론 및 아쉬운 점

**1. 주제 및 목표 소개**

“창업 지원 전문 스타트업 기업 창업을 위한 상권 분석”

일반적이지 않은, 흥미로운 주제입니다. 국가 지표 체계에서 2021년에 집계한 대한민국의 창업률 변화 그래프(그림 1)을 보면, 2013년 이후로 대한민국의 창업률은 13.9%에서 꾸준히 증가하여 2020년에는 15.5%를 기록하였습니다. 창업률은 갈수록 증가하는 추세이나, 대한민국은 기업의 소멸률이 높고 생존률도 낮아 위험 부담이 크게 작용합니다. 따라서, 성공적인 창업을 위해서는 철저한 분석과 준비가 요구될 것입니다. 저희 조에서는 이런 틈새 시장을 공략하여 창업 전문 기업을 설립한다는 목표를 설정하고 주제를 선정하였습니다. 이용자의 성공적인 창업을 지원하기 위해 상권의 특징을 분석하고, 어떤 업종을 창업하는 것이 유리할지 분석해 보고자 합니다.

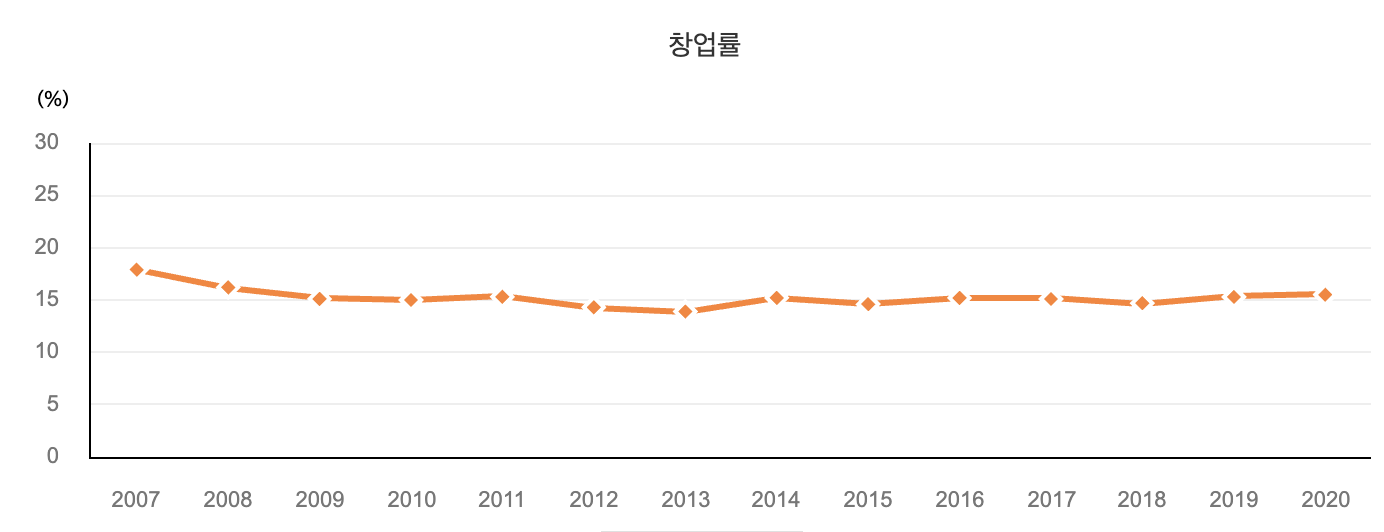


그림1. 대한민국의 창업률 변화 그래프

**2. 시스템 아키텍쳐**

저희 조는 Hive, Hadoop, Jupyter 를 사용하고, Github(<https://github.com/tiger9803/BDP>)를 통해 협업하며 프로젝트를 진행하였습니다.

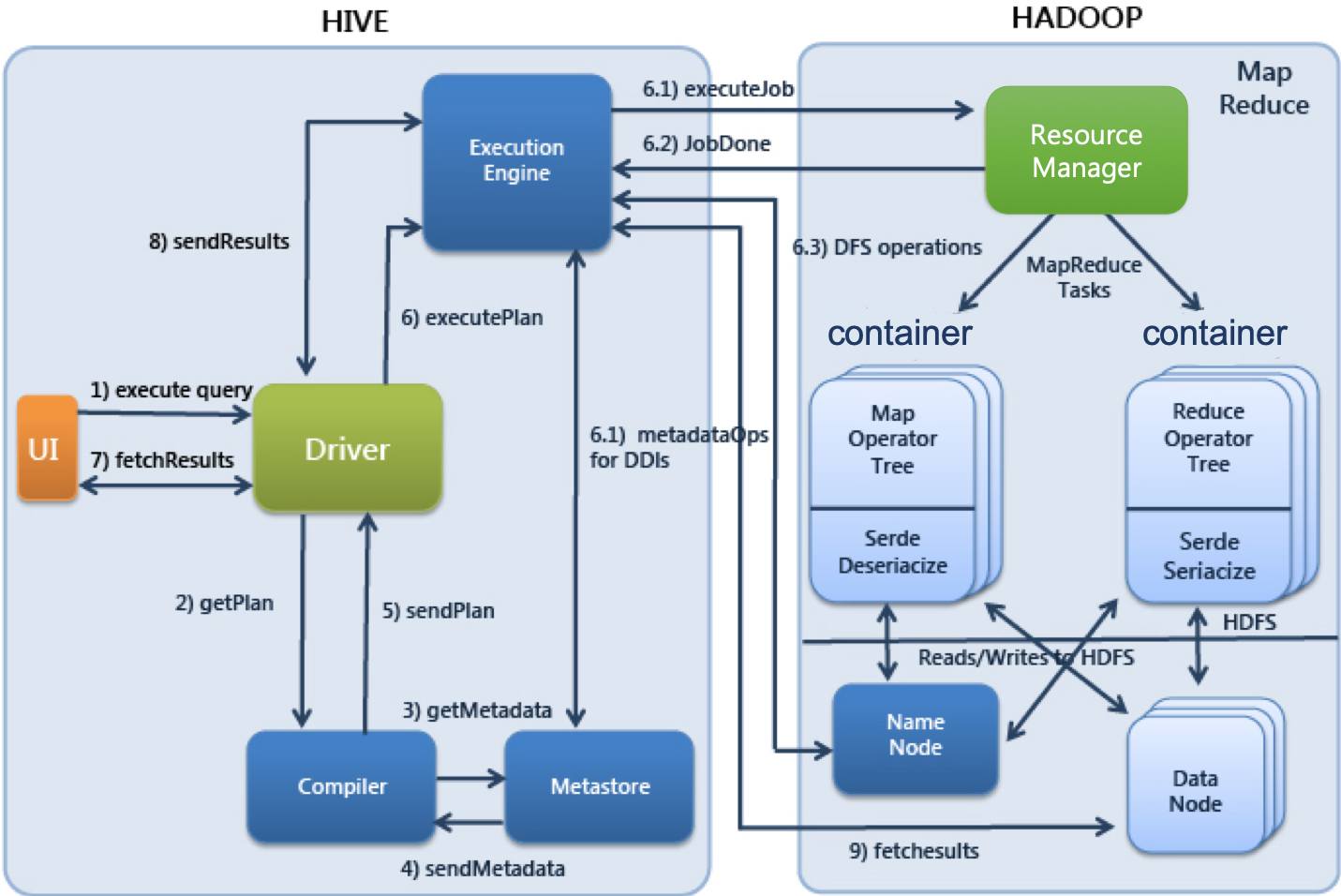


그림2. Hive와 Hadoop간 연동에 대한 설명

Hive에서 UI는 사용자가 쿼리 및 기타 작업을 시스템에 제출하는 사용자 인터페이스입니다. CLI, Beeline, JDBC 등이 있습니다. Driver는 쿼리를 입력받고 작업을 처리합니다. 사용자 세션을 구성하고, JDBC/ODBC 인터페이스 API를 제공합니다. Compiler는 metastore를 참고하여 쿼리 구문을 분석하고 실행계획을 생성합니다. Metastore는 DB, Table, 파티션의 정보를 저장합니다. Excute Engine은 컴파일러에 의해 생성된 실행 계획을 실행합니다.

우선, 사용자가 제출한 SQL문을 driver가 compiler에 요청하여 metastore의 정보를 이용해 처리에 적합한 형태로 컴파일합니다. 컴파일된 SQL을 실행하면, 리소스 매니저가 클러스터의 자원을 적절히 활용하여 실행시켜 줍니다. 실행 중 사용하는 원천데이터는 HDFS등의 저장장치를 이용합니다. 이후 실행결과를 사용자에게 반환시켜 주는 순서로 Hive가 실행됩니다.

Hadoop에서 Hadoop Common은 hadoop의 다른 모듈을 지원하기 위한 공통 컴포넌트 모듈입니다. Hadoop HDFS는 분산저장을 처리하기 위한 모듈입니다. 여러개의 서버를 하나의 서버처럼 묶어서 데이터를 저장합니다. Hadoop YARN은 병렬처리를 위한 클러스터 자원관리 및 스케줄링을 담당합니다. Hadoop Mapreduce는 분산되어 저장된 데이터를 병렬 처리할 수 있게 도와주는 분산 처리 모듈입니다. Hadoop Ozone은 hadoop을 위한 오브젝트 저장소입니다.

저희 조에서 Hive를 사용한 이유는 데이터베이스에 직접 질의하는 것에 비해 부하가 덜하고 빠르게 처리할 수 있다는 장점 때문입니다. Hive를 통해 추출한 데이터는 hadoop과 jupyter를 통해 분석을 진행하였습니다.

**3. 데이터 분석 과정**

저희 조는 “서울 열린데이터 광장”에서 서울시 우리마을가게 상권분석 서비스 데이터를 다운로드 받았습니다. 여러가지 세부 데이터들이 존재했지만, 상권별 직장인구 수 데이터와 상권별 추정매출, 상권별 점포 현황 데이터를 선택하여 사용하였습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

그림 3. 상권별 직장인구 수

상권별 직장인구 수 데이터는 총 43,232개의 row로 이루어져 있으며, 총 직장인구 수, 남성 및 여성 직장인구 수, 연령대 별 직장인구 수 등의 column이 존재하는 데이터입니다.

테이블이(가) 표시된 사진

자동 생성된 설명

그림 4. 상권별 추정 매출(2021)

상권별 추정 매출 데이터는 총 128,703개의 row로 이루어져 있으며, 분기당 매출 금액, 분기당 매출 건수, 주중, 주말 및 요일별 매출 비율 등의 column이 존재하는 데이터입니다.

테이블이(가) 표시된 사진

자동 생성된 설명

그림 5. 상권별 점포 현황(2021)

상권별 점포 현황 데이터는 총 299,164개의 row로 이루어져 있으며, 상권별 업종 및 점포 현황, 개업률, 폐업률 등의 column이 존재하는 데이터입니다.

데이터 전처리 과정에서 상권별 추정 매출과 상권별 점포 현황 데이터는 이전 년도 자료를 모두 포함할 경우 데이터가 너무 방대하여 2021년 데이터만을 사용하게 되었습니다. 그에 맞춰 상권별 직장인구 수 데이터에서도 2021년에 해당하는 row만 따로 뽑아내어 사용하게 되었습니다. 또한, 인구가 적은 곳의 데이터는 큰 의미가 없다고 판단하여, 1,000명 이상인 곳의 데이터만 뽑아 사용하게 되었습니다. 또한 요식업과 관련된 점포에 대해서만 분석을 진행하였습니다.

저희가 데이터 분석 과정에서 확인하고자 하는 내용은 다음과 같습니다.

1. 매출이 높으면 폐업률이 낮은가

2. 총 직장인구 수가 높으면 폐업률이 낮은가

그리고 추가적으로, 요식업에 국한된 것이 아니라 모든 업종 중 어떤 업종이 유동인구가 많은 곳에서 유리하고 불리한지를 분석해 보고자 하였습니다.

데이터 전처리는 Hive에서 쿼리를 작성하여 진행하였습니다. 앞서 소개한 3개의 테이블들을 필요한 row와 column만 추출한 뒤 적절하게 join을 진행하였습니다. 그 후 유동인구가 1,000명 이상인 곳의 업종별 폐업률을 추출하는 쿼리를 작성하여(그림 6) 도출된 결과를 확인하였습니다.(그림 7)

다음으로는, 유동인구가 1,000명 이상인 곳의 업종별 평균 매출과 인구 수를 추출하는 쿼리를 작성하여(그림 8) 도출된 결과를 확인하였습니다.(그림 9)

마지막으로, 위의 두 과정을 합쳐 유동인구가 1,000명 이상인 곳의 업종별 평균 폐업률, 매출, 인구 수를 모두 추출하는 쿼리를 작성하고(그림 10) 도출된 결과(그림 11)를 가지고 분석을 진행하였습니다.

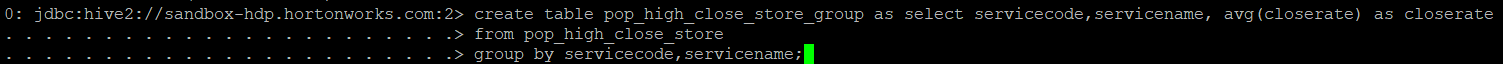


그림 6.

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 7.

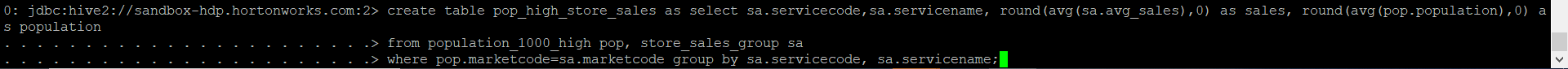


그림 8.

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 9.

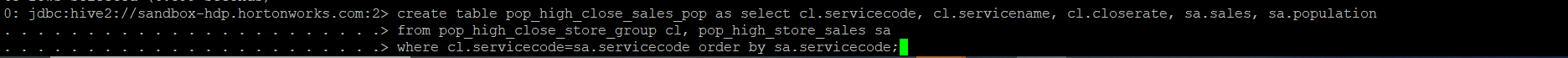


그림 10.

텍스트, 화면, 스크린샷이(가) 표시된 사진

자동 생성된 설명

그림 11.

데이터 분석은 hadoop을 사용했습니다. 우선, 매출과 폐업률의 상관관계를 알아보기 위해 피어슨 상관분석을 진행하였습니다.(그림 12) P-value < 0.05이기 때문에 -0.34의 음의 선형 관계를 가진다는 것을 알 수 있었습니다. 추가로 jupyter에서 시각화를 진행하였습니다. (그림 13)

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 12

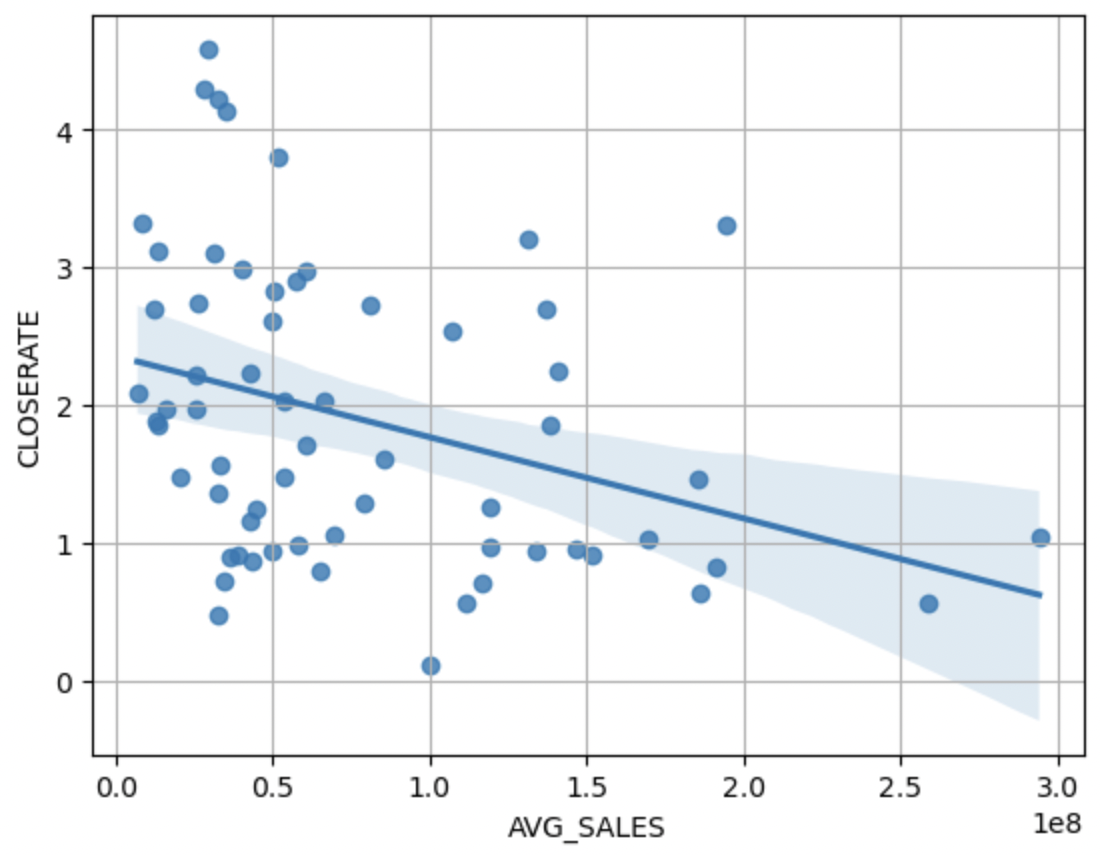


그림 13

다음으로, 테이블을 불러오고(그림 14) 요식업과 관련된 업종만 분리한 뒤(그림 15) 폐업률 순으로 정렬하였습니다.(그림 16) 그 결과 사진에서 보이는 순서로 폐업률이 낮다, 즉 매출이 높다는 결과를 얻게 되었습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 14

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 15

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 16

그 다음, 유동인구와 폐업률의 상관관계를 알아보기 위해 피어슨 상관분석을 진행하였습니다. 마찬가지로 p-value < 0.05이기 때문에 -0.36의 음의 상관관계를 가진다는 사실을 알 수 있었습니다.(그림 17) 시각화도 마찬가지로 진행하였습니다.(그림 18)

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 17

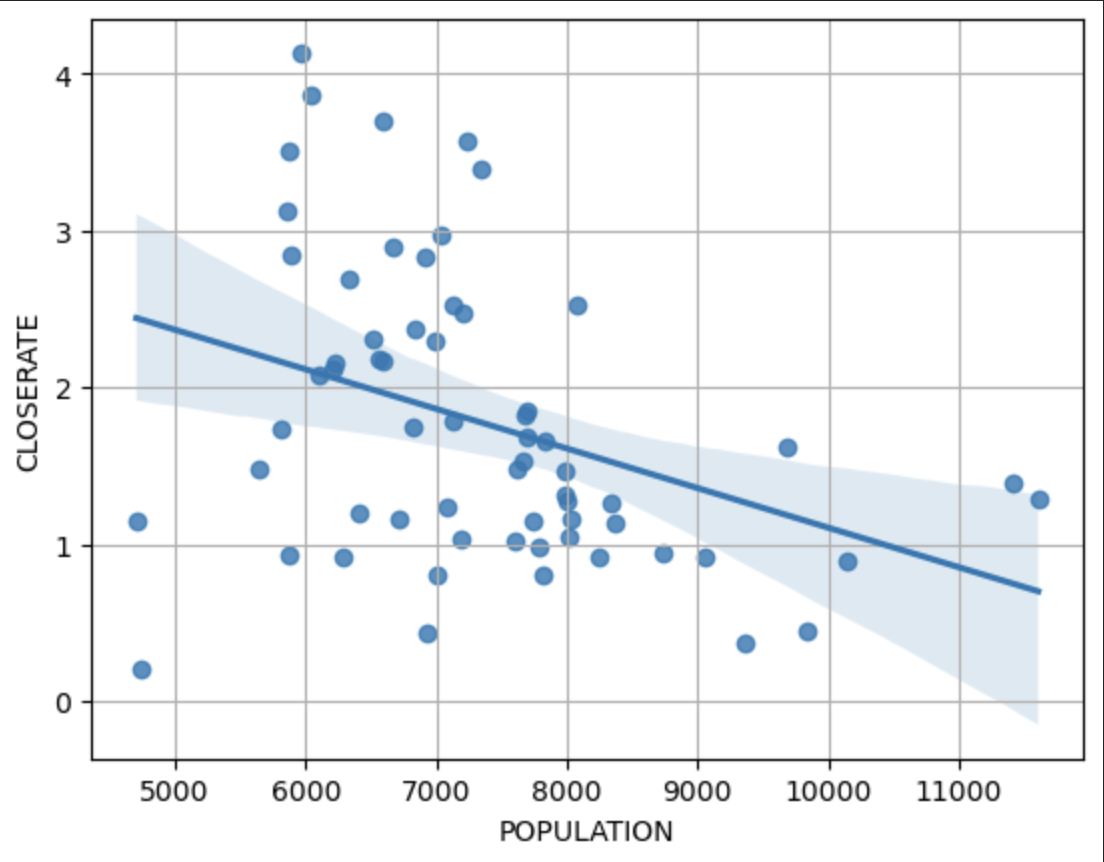


그림 18

추가로, 모든 업종에 대해 유동 인구가 많은 곳에서는 어떤 분야가 유리한지 분석을 진행하였습니다.(그림 19) 요식업 분야는 전반적으로 폐업률이 높고, 가전제품 관련 분야는 폐업률이 낮은 것으로 확인되었습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 19

**4. 결론 및 아쉬운 점**

저희 조는 이번 분석을 통해 매출과 유동인구 수가 높으면 폐업률이 낮아진다는 것을 알게 되었습니다. 그러므로 매출이 높은 업종을 선택하고, 유동인구 수가 많은 곳에 창업을 하면 성공할 확률이 높아집니다. 그러나 이는 너무나 당연하게 알려져 있는 사실입니다. 추가적인 분석을 통해 이용자 맞춤 창업 지원 서비스를 제공할 수 있었다면 더 좋은 분석이 되었을 것이라 생각합니다. 다음으로, 요식업 분야에서 어떤 업종이 폐업률이 낮은지 확인할 때, 매출과 폐업률은 음의 상관관계가 있다는 사실을 가져와 적용하였습니다. 그러나 매출과 폐업률의 상관관계는 요식업 분야에 국한된 것이 아닌 모든 분야에 적용되는 것이었기 때문에, 결과를 도출할 때 조금은 잘못된 방향으로 흘러갔을 수도 있겠다는 생각이 들었습니다.