빅데이터 분석기획

- 도메인 이슈 도출하기

(1) 인터뷰를 통한 조사
인터뷰를 위한 방식으로 Top Down, Bottom Up 방식 중 선택하여 인터뷰를 실행하여 조사한다.
(가) 인터뷰 질의서 항목 구성한다.
인터뷰 질의서 구성 시 필수항목으로 대상업무 파악 항목, 분석주제 이해를 위한 항목, 대상업무 질의서에 대한 원인파악 항목, 추가조사 및 관련 조사 질의 항목을 구성한다.
(나) 인터뷰 내용분석
인터뷰 일정에 따른 인터뷰 수행 후 특정 주제별로 정리하면 문제파악, 개선방향
도출 등을 정리한다.

1) 대상 분야 업무 질의서 구분한다.
인터뷰 진행 후 대상업무 분야별로 질의서를 구분하고 원인을 정리한다.
2) 이슈(issue)별 대상 시스템 및 데이터를 확인한다.
이슈별 대상 시스템 및 데이터별 종류, 특성 등을 목록으로 정리한다.
3) 현행 분석방식을 확인한다.
현재 분석방식 및 주기를 확인하고 이슈로 구분된 내용과 현재의 분석환경과의
제약점 여부를 추출한다.
4) 추가 인터뷰를 통한 개선방향을 정리한다.
1차 인터뷰 결과 정리 후 개선 가능한 분석접근 방안을 제시하고 현업들로부터
개선 의견을 청취하여 정리한다.

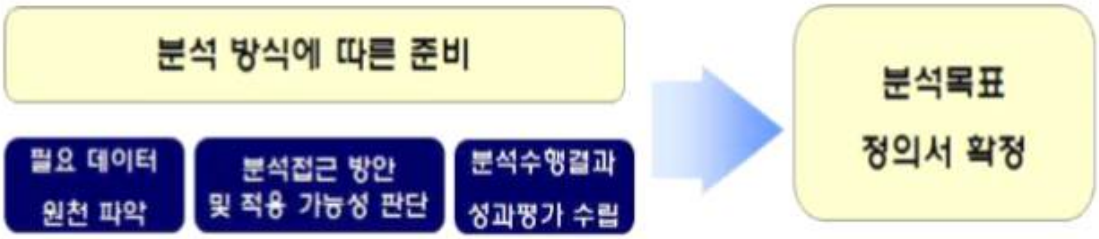
인터뷰 진행 시 질문 예시
① 귀하께서 맡고 계시는 업무와 업무 내의 역할은 어떤 것입니까?
② 귀하의 팀에서 운영 또는 관리하는 시스템의 종류는 어떤 것이 있습니까?
③ 관련 현업조직(고객사)의 조직구성은 어떻게 되어 있습니까?
④ 현업조직의 업무에 대해 간략하게 설명하여 주십시오.
⑤ 현업조직의 업무가 현재 XXX 시스템을 통해 관리되고 있습니까? 관리되고 있다면 관련된 프로세스는 몇 개나 있습니까?

⑥ 귀하께서 맡고 계시는 기능업무 중 업무적 이슈가 있다면 어떤 것이 있는지 구체적으로 말씀해 주십시오.
⑦ 귀하께서 언급하신 이슈에 관한 내용이 어떠한 원인에 의해 발생하고 있다고 생각되는지 말씀해 주십시오.
⑧ 귀하께서 언급하신 이슈에 관한 내용이 업무 상황의 개선으로 해결될 문제인지 신규 시스템 도입 또는 기존 시스템의 개선으로 해결 가능한 것인지 구분하여 주십시오.
⑨ 현업조직의 업무 중 선/후행 연결된 업무나 관련된 업무가 있다면 어떤 것이 있는지 말씀해 주십시오.
⑩ 선/후행 업무처리 중 이슈가 있다면 어떤 것이 있는지 말씀해 주십시오.
⑪ 추후 업무와 관련하여 시스템 개발 또는 개선 계획을 하고 있다면 구체적으로 말씀해 주십시오.
⑫ 현행업무 중 현업에 제공하시는 분석 리포트(또는 일련의 정보)의 종류가 어떤 것이 있는지 말씀해 주십시오.
⑬ 제공되는 분석 리포트를 생성하기 위해 소요되는 시간과 보고주기는 어떻게 되는지 말씀해 주십시오.
⑭ 제공되는 분석 리포트의 주요 내용은 어떤 것입니까?
⑮ 분석 리포트를 통해서 현업에서 얻으려는 가장 큰 목적은 무엇이라 생각하십니까?
⑯ 제공되는 분석 리포트의 시기가 현재보다 앞당겨질 수 있다면 현재보다 효 과적일 수 있습니까?
⑰ 관련 업무에서의 특별한 사고/사건이 있었습니까? (현업중심에서) 있었다면 업무나 회사에 미치는 영향은 무엇이었습니까?
⑱ 사고/사건이 있었다면 주요 원인은 무엇이었습니까?
⑲ 사고/사건에 대해 미리 관련 정보를 받았으면 방지할 수 있었다고 보십니까?
⑳ 현재 현업 또는 시스템 담당자로서 어떤 데이터 분석이 필요하다고 생각하 십니까?
㉑ 사고/사건과 별도로 현업에서 자주 요구하거나 개선하고자 하는 업무 이슈가 있습니까?
㉒ 현업에서 개선하고자 하는 부분에 대해 구체적으로 실행하고자 하는 계획을 세우고 있습니까?
㉓ 자신이 맡은 업무 외에 연관된 업무와 연관된 이슈가 있으면 말씀해 주십시 오. (예: 이 부분의 일이 늦어지거나 정보가 자주 오류가 발생하여 업무에 방해를 많이 받거나 질책을 받고 있다고 생각되는 부분)

(2) 인터뷰 결과 정리한다.
인터뷰 항목별 내용을 현재의 이슈, 원인, 관련 시스템, 개선 기대사항 등으로 구분되게 정리 후 필요에 따른 추가적인 조사자료 요청과 인터뷰를 진행하여 최종적으로는 문서로 결과를 정리한다.
정리된 대상 업무 이슈들에 대해 분석관점에서 어떤 해결방안이 있을 수 있는지 고민 후 해당 해결방안이 유효한지를 판단하는 것이 개선과제 작성에서 중요하다. 인터뷰 결과서의 경우 1차, 2차로 구분하여 작성하고 문서화해야 한다.

- **분석목표수립** : 도메인 이슈(issue) 정하기에서 정리된 내용을 바탕으로 개선방향에 알맞은 분석목표를 수립해야 한다. 분석방법에 따라 필요한 데이터 원천, 데이터 입수 난이도, 분

석방법 및 개별 분석의 난이도, 분석 수행주기, 분석결과에 대한 성과평가 기준을 고려하여 목표 수준을 정리해야 하며, 현시점에 현실적으로 가능하게 분석목표를 수립해야 한다.



(1) 필요 데이터 원천을 파악한다.
도메인 이슈 정하기에서 선별한 가능성 있는 분석주제와 대상 시스템을 대상으로 개별 테이블 정보 및 메타정보를 조사한다.
(2) 분석 접근 방안 및 적용 가능성을 판단한다.
분석 목표에 부합되는 분석접근 방안을 목록화한다. 현재의 분석환경, 데이터 환경과 앞으로 구축하고자 하는 분석환경을 고려하여 적용 가능성에 대한 의견을 기술한다. 기술적인 이슈로 적용 가능성이 낮을 경우는 분석 목표를 고려한 대안을 고민하고 제시해야 한다.
(3) 분석 수행결과에 대한 성과평가 기준을 마련한다.
분석 수행결과에 대한 성과평가 기준은 정성적인 측면과 정량적인 측면으로 구분해야 한다. 현 단계에서는 분석 수행 이전이므로 성과평가 기준 수립 후 데이터 가설/검증 단계, 데이터 분석 모형 검증 단계, 실제 운영 단계 등으로 구분하여 측정하도록 한다.
분석 목표 정의서를 확정한다.
분석 방식에 따른 준비과정을 통해 정리된 내용을 바탕으로 정의서를 작성한다. 작성된 정의서에 대하여 실무자들과의 워크숍을 통하여 확정한다. 피드백 내용이 있을 경우는 별도로 수정 후 확정하도록 한다.
<분석목표 예시서>

분석기본 정보	분석 명칭	시장품질 및 생산품질 한계불량률 사전 예측	분석 목표 확정일	2016년 6월 10일
	분석 목적	사전 예측 통한 제품 불량률 감소	분석 목표 워크숍	2016년 6월 8일
	분석 우선순위	상	Owner 조직명	시장품질팀
	분석접근 방안	- 시장품질 불량이슈는 생산품질단계에서 사전 발생하므로 부품별 이슈파악 후 예측분석을 통해 부품별 불량률을 예측하여 사전조치가 가능하게 하는 것으로 목표임 - 부품 불량 예측 시 민원 콜유형 내용과 부품 TEST 내용, 날씨 정보 활용		
성과측정	정성적 기준	- 신규기법/기술 : 다양한 예측기법 및 기계학습 활용 - 외부데이터 : 기상청 날씨 데이터 활용 - 신규 데이터 : 품질 VOC 데이터 활용 - 기타 : 국내 품질학회 경진대회 참여		시장품질팀 합의
	정량적 기준	- 기존품질 분석대비 3% 이상 불량률 개선 - 기존민원대비 10% 이상 감소 목표 - 분석가설 : 20개 수립 - 데이터 모형 검증 : 80% 정확도 입증		시장품질팀 합의
데이터 정보	내부 데이터	- 시장품질 모니터링 정보 (대상테이XX외 7개) - 생산품질 측정정보 (대상테이블XX외 5개) - 콜센터 민원정보 (대상테이블XX외 8개)	데이터입수 난이도	중
	외부 데이터	- 기상청 날씨 정보 (2014~2016년 4월분)	데이터입수 난이도	하
데이터 분석 적용성 판단		- 실시간 분석은 분석 인프라 문제로 이번에는 제외 함 - 민원콜분석을 위한 텍스트 분석은 중요도를 감안하여 도입 후 진행하기로 함		

* 분석 목표 수립 시 기획의도를 명확히 하고 성과측 정 방안에 대해 관련 조직간 이견이 없도록 하는 것이 중요하다.

- **프로젝트계획 설계** :데이터 분석 프로젝트를 위한 계획수립을 위해 책정된 소요비용에 관련된 예산, 예상되는 소요기간, 현재의 IT 환경(데이터 분석 솔루션, 플랫폼, 데이터 현황)을 고려해야 한다. 이를 위해서는 앞선 단계에서 만들어진 분석 목표 정의서 내용을 확인하고 해당 데이터 분석 프로젝트의 주관부서와의 협의를 통하여 최종적인 프로젝트 계획을 수립해야 한다.

<인건비 배분기준>

인원 구분	투입 기간 및 인원별 단가
프로젝트 매니저	• 투입 기간: X개월 • 인원별 단가:(단가등급 초/중/고, 개별 단가 표시)
데이터 분석가	• 투입 기간: X개월 • 인원별 단가:(단가등급 초/중/고, 개발 단가 표시) • 여러 명인 경우 투입 월을 기준으로 개별적으로 표시
데이터 처리 엔지니어	• 투입 기간: X개월 • 인원별 단가:(단가등급 초/중/고, 개별 단가 표시) • 여러 명인 경우 투입 월을 기준으로 개별적으로 표시
기타 인원	• 투입목적별 기재 • 투입 기간: X개월 • 인원별 단가:(단가등급 초/중/고, 개별 단가 표시) • 투입목적별 인원이 여러 명인 경우 개별적으로 표시

<프로젝트 필수 산출물 정의>

산출물 구분	산출물 구분
데이터 분석과제 계획서	• 데이터 분석 목표 정의서 포함 기재 • 프로젝트 일정계획, 자원배분 계획, 의사소통 계획 포함
데이터 탐색 보고서	• 데이터 수집대상 내용 포함 기재 • 데이터 후보 변수 도출과정 및 최종 후보 변수 목록 포함 • 데이터 분석가설별 유의성 검증 내용 포함
데이터 모델링 및 검증 보고서	• 데이터 모델링 방안 및 실험계획 포함 기재 • 데이터 모델링 개발 스크립트 • 데이터 모형 비교검증 위한 검증내용 포함
기타 보고서	• 별도의 화면 개발 시 화면설계 등에 대한 산출물 • 데이터 분석 모델 유지보수 위한 산출물 • 데이터 분석 모델 교육을 위한 산출물 • 기타 필요 산출물 정의

- **보유 데이터 자산 확인** : 정의된 분석목표와 프로젝트 계획에 따라 데이터 분석 시작 전 현재 보유하고 있는 데이터를 점검해야 한다. 분석목표 또는 서비스 기획방향에 따라 보유 데이터 중 어떤 시스템의 데이터, 보관 데이터 중 어느 주제영역의 데이터 등을 활용할지 사전에 파악할 수 있으므로 해당 데이터의 상황을 정리할 수 있다. 데이터의 품질, 분량, 수집경로 및 데이터의 유형 등이 정리되면 데이터 분석 단계에서의 데이터 파악 및 전처리 과정의 시간을 많이 절약할 수 있다.

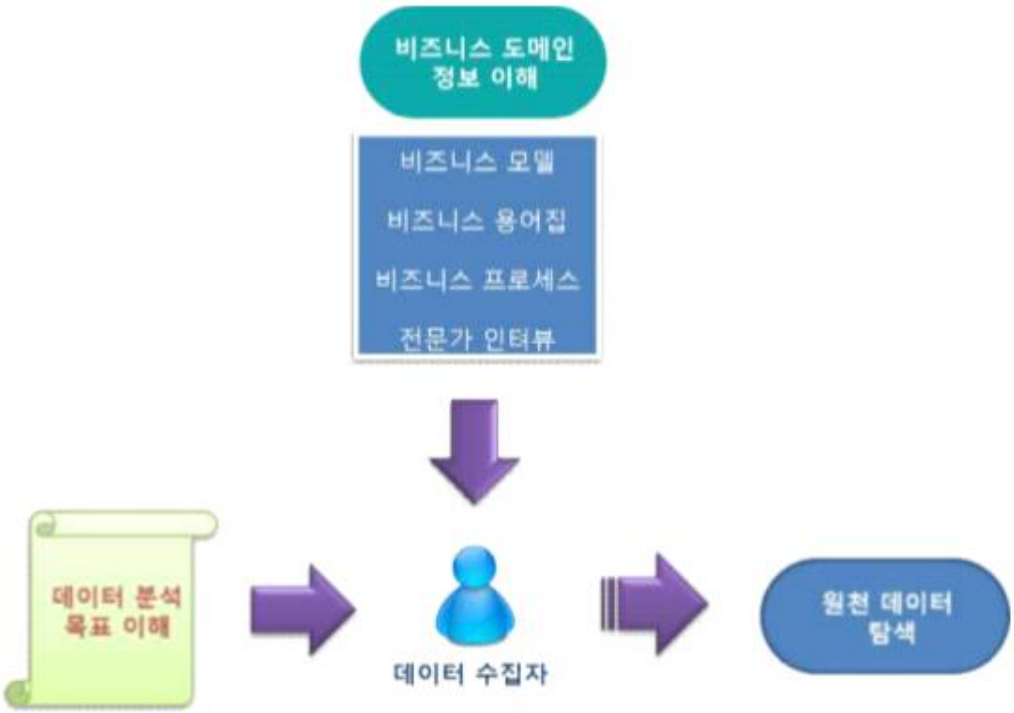
점검항목	점검내용
데이터 분량 확인	• 테이블 내 필요 칼럼별 확인 • 칼럼별 데이터 축적 기간 및 분량 확인
데이터 완전성	• 데이터 내 필요한 대상과 속성을 포함하는지 판단 • 데이터 누락 또는 결측값의 비율 확인
데이터 일관성	• 데이터 속성 간 관계 확인 • 데이터 상위/하위 간 관계에서의 값의 일치성 확인 • 데이터 유형과 값의 일치성 확인
데이터 정확성	• 데이터의 편향(bias)과 분산 확인 - 편향성이 큰 경우는 측정값이 지속적인 영향 받는 경우임 - 분산이 큰 경우는 표본의 대표성이 낮을 수 있음

<보유데이터 자산 확인하기 프로세스>

2. 데이터 수집

데이터 수집을 위해서 데이터 수집자는 데이터 분석 목표를 이해하고, 비즈니스 도메인에 대한 이해를 바탕으로 원천 데이터를 탐색해야 한다.

비즈니스 도메인의 이해를 통한 원천 데이터 탐색프로세스



3. 빅데이터 저장

빅데이터 저장시스템이란 대용량 데이터 집합을 저장하고 관리하는 시스템으로 사용자에게 데이터 제공 신뢰성과 가용성을 보장하는 시스템을 말한다. 빅데이터 저장시스템을 구성하기 위해서는 하드웨어인 (분산)저장시스템과 데이터베이스 기술에 기반을 둔 다양한 형태의 빅데이터 저장 소프트웨어를 선택하여 구성하여야 하나, 이 장에서는 빅데이터 관련 저장 기술 및 제품 관련 즉 소프트웨어 부문에 집중하도록 한다.

< 빅데이터 저장시스템 준비상황 체크리스트>

구축 단계 고려요소	체크 항목	준비 상황
1. 빅데이터 저장시스템 도입 및 환경 설정	- 시스템 설치를 위한 하드웨어 환경 . 설치 장소, 전력량, 항온항습장치 . 네트워크 연결 . 서버 사양 및 필요 대수	
	- 소프트웨어 환경 . 설치 서버의 필요 SW: OS 등 확보 . 분산 환경 설치 여부 . 시스템 모니터링 환경 설치 여부	
2. 시스템 구축 일정 및 역할	- 빅데이터 모델 설계 일정 및 타당성	
	- 저장시스템 설치/설정 및 테스트 일정	
	- 샘플 데이터 적재 및 테스트 일정	
	- 실데이터 적재 및 검증 일정	
	- 가동 일정 및 역할	
3. 빅데이터 수집 및 처리 시스템과 연계 방안	- 수집 시스템과 연계 방안 수립(일정, 역할)	
	- 처리 분석 시스템과 연계 방안 수립(일정, 역할)	
4. 빅데이터 저장시스템 검증 및 운영 방안	- 저장시스템 최종점검 계획/운영 조직 체계 수립	

- 빅데이터 처리

빅데이터 처리시스템이란 대용량 데이터를 분산 병렬 처리하고 관리하는 시스템으로 사용자에게 데이터의 유형에 따라서 실시간 처리나 배치(batch) 처리를 할 수 있도록 하는 프레임워크를 제공한다. 빅데이터 처리시스템은 대규모 양의 데이터의 수집/관리/유통/분석을 처리하는 일련의 분산 병렬처리 프레임워크를 말한다. 빅데이터 처리를 통해서 유용한 정보 및 데이터 내면에 있는 지식을 찾아낼 수 있다. 이런 정보를 찾기 위한 데이터 가공 및 분석 과정 전반을 지원하는 것이 빅데이터의 처리 기술이다. 현재의 대규모 데이터 처리를 위한 확장성, 데이터 생성 및 처리 속도를 지원하기 위한 실시간 처리, 비정형 데이터 처리에 대한 지원 등이 빅데이터 처리시스템의 필수 요소이다.

- 분석S/W

(1) 하둡(Hadoop)

오픈 소스인 아파치 하둡은 하둡 분산 파일시스템(HDFS)과 맵리듀스를 구현한 빅데이터 처리 프레임 워크이다. 하둡은 적은 비용으로 큰 데이터의 처리가 가능하고 다양한 에코 시스템으로 구성되어 있다. 하둡의 파일시스템은 파일을 블록 단위로 나누어 저장하고, 결함 허용을 위해서 각 블록의 복사본을 유지한다. 하둡에서는 데이터가 있는 곳에서 처리하고, 일반적인

하드웨어 사양을 사용하기 때문에 하드웨어적인 고장을 고려하여 설계해야 한다. 그래서 파일을 저장 시에 기본적으로 블록으로 나누어 저장할 때 각 블록을 여러 개 중복 저장할 수 있다. 이때 중복하는 블록 수를 하둡에 설정하여 관리할 수 있고, 기본값은 세 개의 중복된 블록을 가지는 것이다. 복사본의 수는 특정 파일에 대해서 다른 복사본 수를 가질 수 있다. 설정에 의하여 블록이 저장될 랙을 지정할 수 있다.

(2) 알(R)

알은 통계계산 및 시각화를 위한 개발환경을 제공하며 이를 통해서 기본적인 통계 및 모델링과 데이터 마이닝 등이 구현할 수 있다. 통계적 언어로 다양한 통계분석과 예측 분석이 가능하다.

(3) 프레스트(Presto)

빅데이터 분석 도구인 프레스트는 페이스북에서 개발된 하둡을 위한 SQL 처리 엔진으로 SQL 언어로 데이터를 빠르게 분석할 수 있다. 클라우데아 임팔라와 아파치 타조 등과 유사하다.

(4) 빅쿼리(BigQuery)

빅쿼리는 구글에서 개발한 대용량 데이터를 처리하는 엔진으로 이를 사용하기 위해서는 먼저 분석할 데이터를 구글 시스템에 업로드하고, 빅쿼리 API를 사용하여 질의를 전송하는 방식이다. 구글 클라우드 스토리지와 함께 이용하며 최대 2TB까지 데이터를 업로드하여 무료로 분석할 수 있다.

(5) 맵리듀스(MapReduce)

맵 함수와 리듀스 함수 기반으로 구성되는 키, 값 리스티를 기반으로 데이터를 병렬처리하는 방식이다.

(6) 서밍버드(Summingbird)

스톰과 하둡을 결합한 스트리밍 맵리듀스 시스템으로 배치 작업과 스트리밍 작업을 모두 요구하는 애플리케이션을 수행하는 데 유용한 서비스이다.

(7) 에스퍼(Esper)

실시간 처리를 위한 인-메모리 기술로 여러 이벤트 소스로부터 발생한 이벤트를 대상으로 의미 있는 데이터를 추출하여 그것에 대응하는 작업을 수행하는 서비스이다.