빅데이터

Big data Framework

프레임워크

01

Big data Framework

빅데이터 처리 개요

02

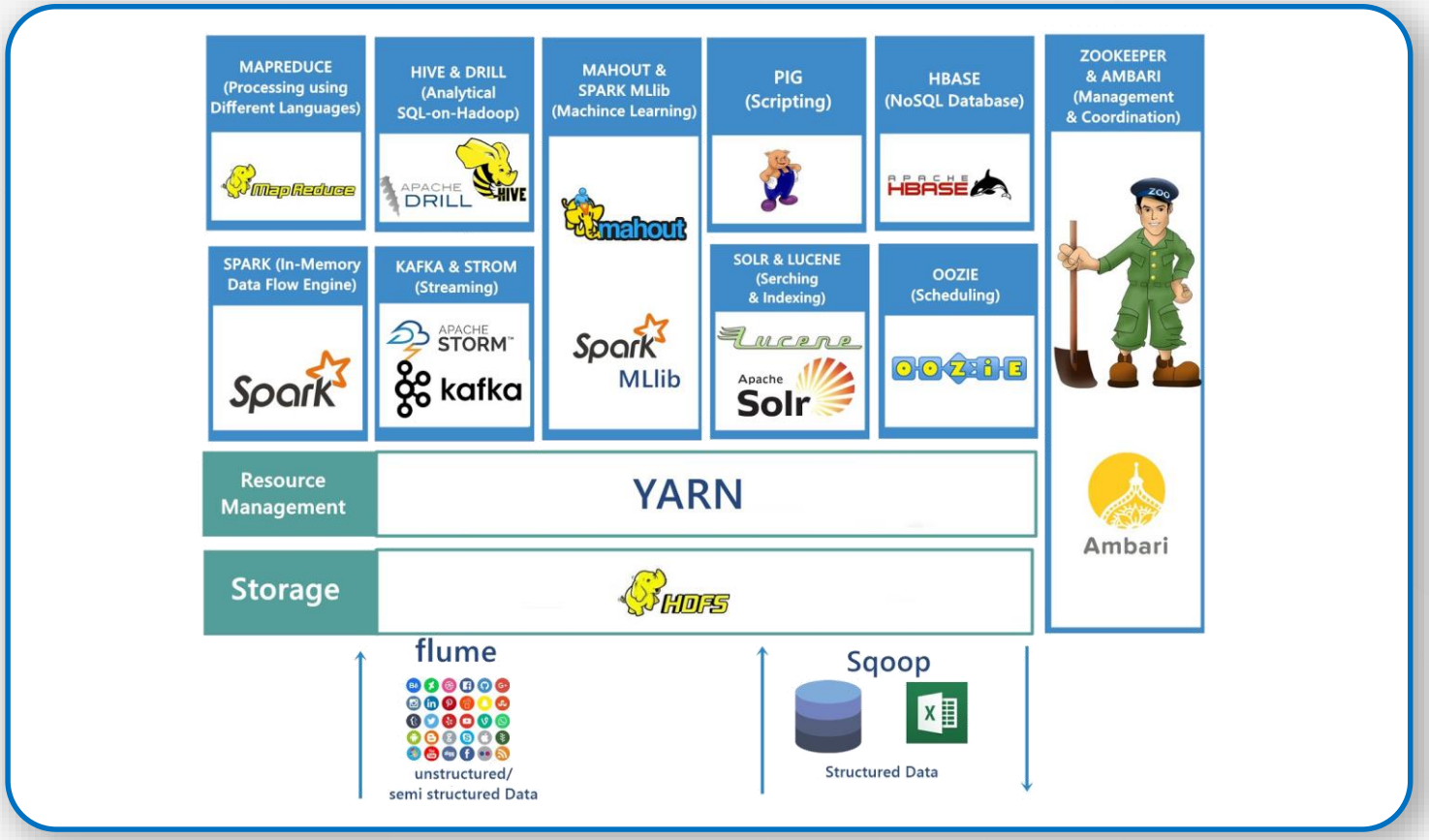
빅데이터 프레임워크의 개념

02 빅데이터 프레임워크의 개념

- 빅데이터 프레임워크의 개념에 대해 설명할 수 있다.
- 빅데이의 프레임워크의 기능에 대해 분석할 수 있다.

1 빅데이터 프레임워크란?

- 빅데이터 수집, 저장, 처리, 분석, 시각화를 손쉽게 가능하도록 라이브러리 및 모듈로 기반 기술을 제공하는 소프트웨어



[출처] http://www.bigintellects.com/2018/06/hadoop-ecosystem.html

1 빅데이터 프레임워크란?

빅데이터 수집, 저장, 처리, 분석, 시각화를 손쉽게 가능하도록 라이브러리 및 모듈로 기반 기술을 제공하는 소프트웨어

과정	영역	개요
생성	내부데이터	데이터베이스(Database), 파일관리시스템(File Management System)
	외부 데이터	인터넷으로 연결된 파일, 멀티미디어, 스트림
수집	크롤링(Crawling)	검색 엔진의 로봇을 사용한 데이터 수집
	ETL (Extraction, Transformation, Loading)	소스 데이터의 추출·전송·변환·적제
저장	NoSQL 데이터베이스	비정형 데이터 관리
	스토리지(Storage)	빅데이터 저장
	서버(Server)	초경량 서버

(1/2)

[출처] P. Warden, "Big Data Glossary". 2011 O'Reilly Media

1 빅데이터 프레임워크란?

- 빅데이터 수집, 저장, 처리, 분석, 시각화를 손쉽게 가능하도록 라이브러리 및 모듈로 기반 기술을 제공하는 소프트웨어

과정	영역	개요
처리	맵리듀스(MapReduce)	데이터 추출
	프로세싱(Proccessing)	다중 업무 처리
분석	NLP (Neuro Linguistic Programming)	자연어 처리
	기계 학습(Machine Learning)	기계 학습으로 데이터의 패턴 발견
	직렬화(Serialization)	데이터 간의 순서화
표현	가시화(Visualization)	데이터를 도표나 그래픽적으로 표현
	획득(Acquisition)	데이터의 획득 및 재해석

(2/2)

[출처] P. Warden, "Big Data Glossary". 2011 O'Reilly Media

2 빅데이터 수집

- 정형, 비정형, 반정형 데이터
- 실시간, 배치 데이터

빅데이터 자동 수집 방법

방법	개요
로그 수집기	▪ 내부에 있는 웹 서버의 로그를 수집 - 웹 로그, 트랜잭션 로그, 클릭 로그, DB의 로그 데이터 등 수집
크롤링	▪ 주로 웹 로봇으로 거미줄처럼 얽혀 있는 인터넷 링크를 따라다니며 방문한 웹 사이트의 웹 페이지라든가 소셜 데이터 등 인터넷에 공개되어 있는 데이터 수집
센싱	▪ 각종 센서로 데이터 수집
RSS 리더/오픈 API	▪ 데이터의 생산·공유·참여 환경인 웹2.0을 구현하는 기술 ▪ 필요한 데이터를 프로그래밍으로 수집
ETL (Extraction, Transformation, Loading)	▪ 데이터의 추출, 변환, 적재의 약자 ▪ 다양한 소스 데이터를 취합해 데이터를 추출하고 하나의 공통된 형식으로 변환하여 데이터웨어하우스에 적재하는 과정 지원

[출처] "빅데이터 시대의 데이터 자원 확보와 품질 관리 방안", 한국정보화진흥원, IT & Futrue Stategy

3 빅데이터 저장

- 분산 파일 시스템
- 대용량 데이터 베이스

대용량 데이터를 저장하는 다양한 접근방식

접근 방식	설명	제품
분산 파일 시스템	컴퓨터 네트워크로 공유하는 여러 호스트 컴퓨터 파일에 접근할 수 있는 파일 시스템	GFS(Goole File System), HDFS(Hadoop Distributed File System), 아마존 S3 파일 시스템
NoSQL	데이터 모델을 단순화해서 관계형 데이터 모델과 SQL을 사용하지 않는 모든 DBMS 또는 데이터 저장 장치	Cloudata, Hbase, Cassandra

[출처] "빅데이터 시대의 데이터 자원 확보와 품질 관리 방안", 한국정보화진흥원, IT & Futrue Stategy

4 빅데이터 처리

- 실시간 처리
- 배치 처리

빅데이터 일괄 처리 기술 | 빅데이터 실시간 처리 기술 | 빅데이터 처리 프로그래밍 지원 기술

- 빅데이터를 여러 서버로 분산하여 각 서버에서 나누어 처리하고, 이를 다시 모아서 결과를 정리하는 분산 병렬 기술 방식
- **구글 맵리듀스**, 하둡 맵리듀스, 마이크로소프트 드라이애드(Dryad) 등이 있음

  - 구글에서 분산 컴퓨팅을 지원할 목적으로 제작·발표한 소프트웨어 프레임워크
  - 함수형 프로그래밍에서 일반적으로 사용되는 맵(Map)과 리듀스(Reduce) 함수를 기반으로 주로 구성

[출처] 빅데이터 시대의 데이터 자원 확보와 품질 관리 방안", 한국정보화진흥원, IT & Futrue Stategy

4 빅데이터 처리

- 실시간 처리
- 배치 처리

빅데이터 일괄 처리 기술	**빅데이터 실시간 처리 기술**	빅데이터 처리 프로그래밍 지원 기술

- 스트림 처리 기술로 강화된 스트림 컴퓨팅을 지원하는 IBM의 InfoSphere Streams (인포스피어 스트림즈)
- 분산 환경에서 스트리밍 데이터를 분석할 수 있게 해 주는 트위터의 스톰(Storm)

[출처] 빅데이터 시대의 데이터 자원 확보와 품질 관리 방안", 한국정보화진흥원, IT & Futrue Stategy

4 빅데이터 처리

- 실시간 처리
- 배치 처리

빅데이터 일괄 처리 기술 | 빅데이터 실시간 처리 기술 | **빅데이터 처리 프로그래밍 지원 기술**

- 분산 데이터를 처리하는 프로그래밍 언어인 구글의 소잴(Sawzall)
- 병렬 처리를 하는 고성능 데이터-플로우 언어와 실행 프레임워크인 하둡 Pig

[출처] 빅데이터 시대의 데이터 자원 확보와 품질 관리 방안", 한국정보화진흥원, IT & Futrue Stategy

4 빅데이터 처리

인프라 기술을 포함한 빅데이터와 연계된 기술들

Cassandra (카산드라)	▪ 분산 시스템에서 대용량 데이터를 처리할 수 있도록 설계된 오픈 소스 데이터 베이스 관리 시스템 ▪ 원래 페이스북에서 개발했으며, 지금은 아파치 소프트웨어 재단에서 한 프로젝트로 관리
Hadoop (하둡)	▪ 분산 시스템에서 대용량 데이터 처리 분석을 지원하는 오픈 소스 소프트웨어 프레임워크 ▪ 구글이 개발한 맵리듀스를 오픈 소스로 구현한 결과물 ▪ 야후에서 최초로 개발하였으며, 지금은 아파치 소프트웨어 재단에서 한 프로젝트로 관리 ▪ 주요 구성 요소로는 하둡 분산 파일 시스템인 HDFS, 분산 칼럼 기반 데이터베이스인 Hbase, 분산 컴퓨팅 지원 프레임워크인 맵리듀스 포함

[출처] "빅데이터 분석: CIO를 위한 미래지향적 아키텍처 기술 그리고 로드맵", IDC

4 빅데이터 처리

인프라 기술을 포함한 빅데이터와 연계된 기술들

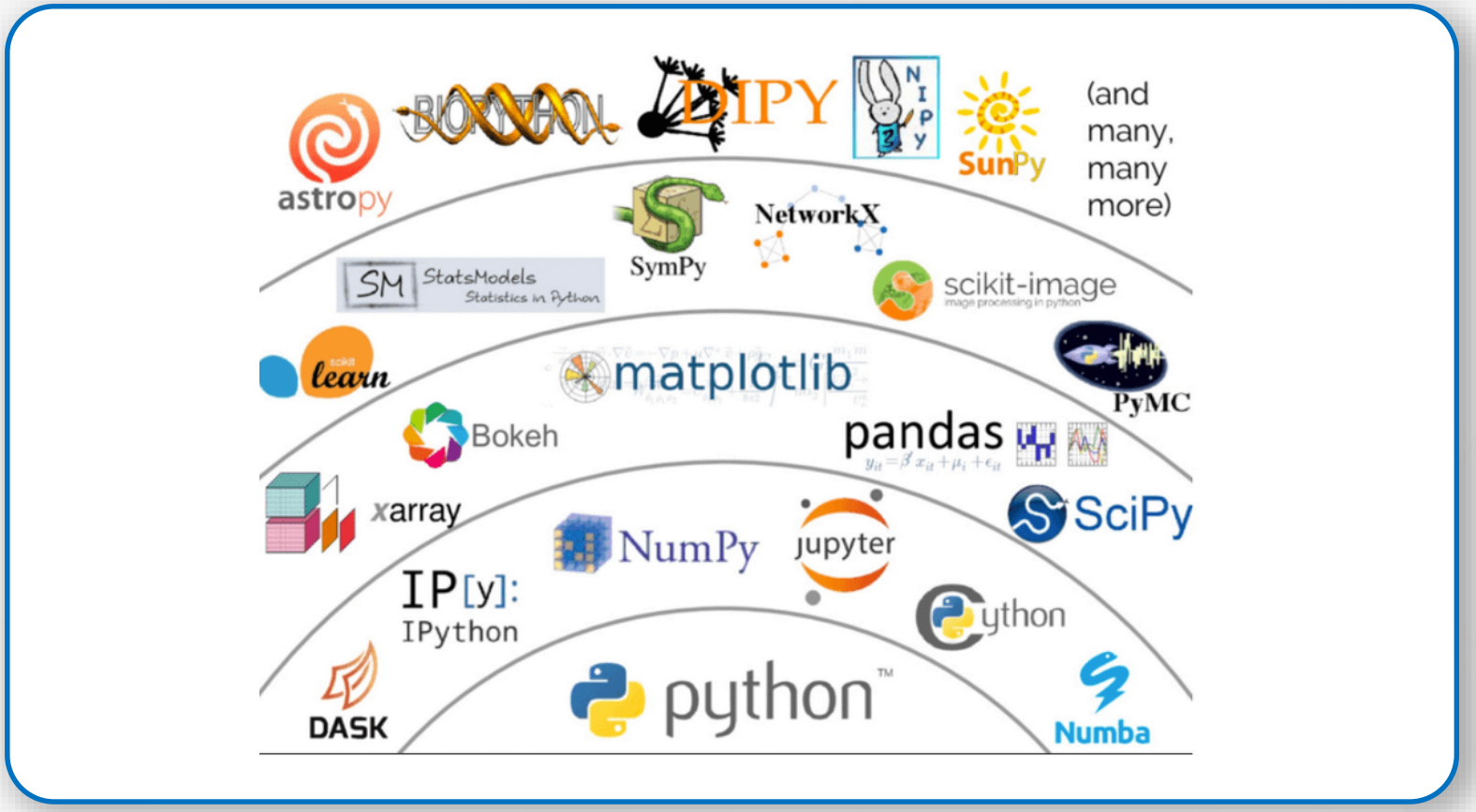
Hbase (H베이스)	▪ 구글의 '빅테이블'을 참고로 개발된 오픈 소스 분산 비관계형 데이터베이스 ▪ 파워셋에서 개발했으며, 현재는 아파치 소프트웨어 재단에서 한 프로젝트로 관리
MapReduce (맵리듀스)	▪ 분산 시스템에서 대용량 데이터를 처리하려고 구글이 제안한 소프트웨어 프레임워크 ▪ 하둡에서도 구현
NoSQL	▪ Not-only SQL 또는 No SQL을 의미 ▪ 전통적인 관계형 데이터베이스와 다르게 설계된 비관계형 데이터베이스 ▪ 대표적인 NoSQL 솔루션으로는 Cassandra, Hbase, Mongo유 등이 있음

(2/2)

[출처] "빅데이터 분석: CIO를 위한 미래지향적 아키텍처 기술 그리고 로드맵", IDC

5 빅데이터 분석

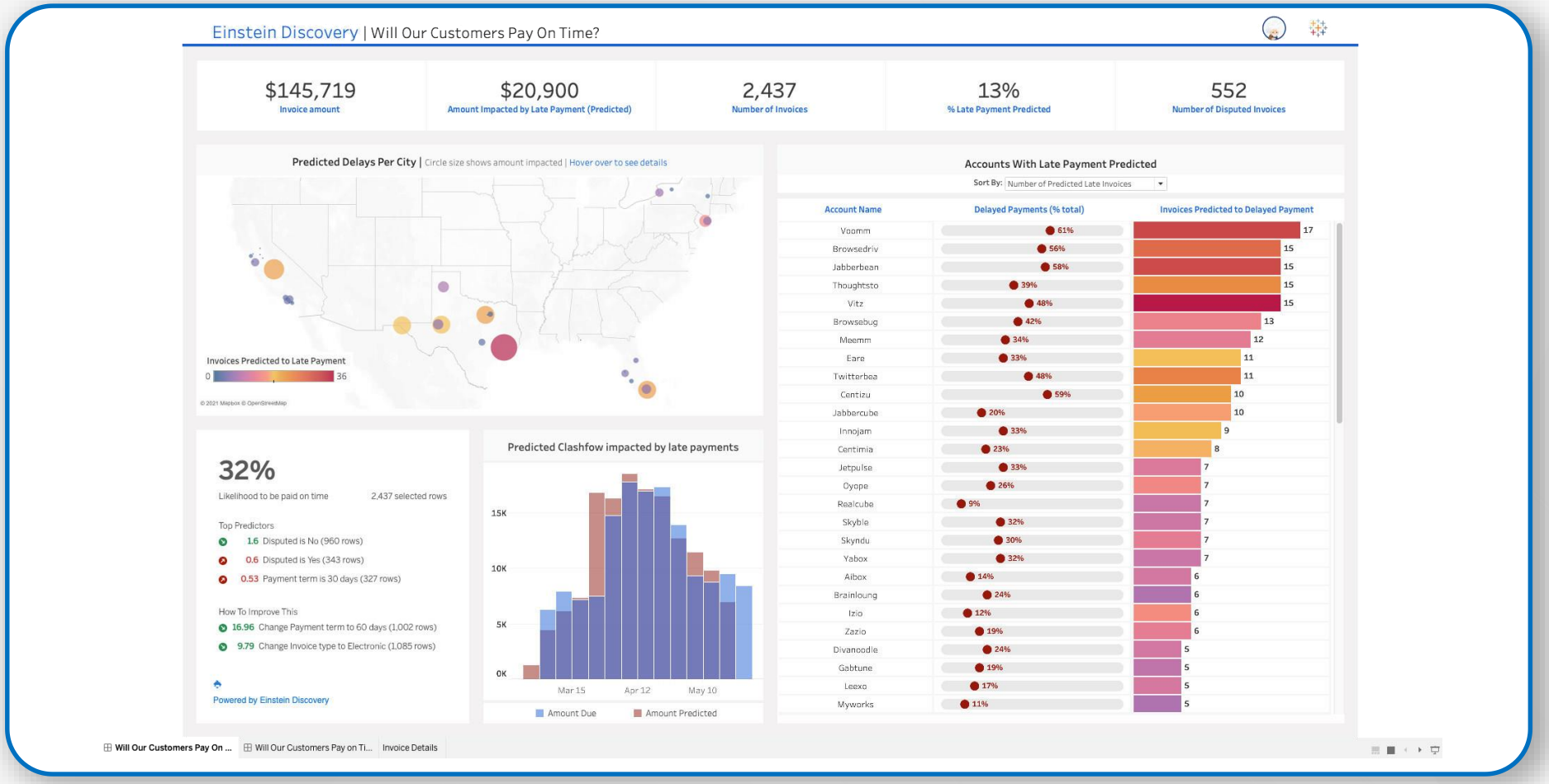
- 통계 기법
- 기계 학습



[출처] https://coiled.io/pydata-dask/

6 빅데이터 시각화

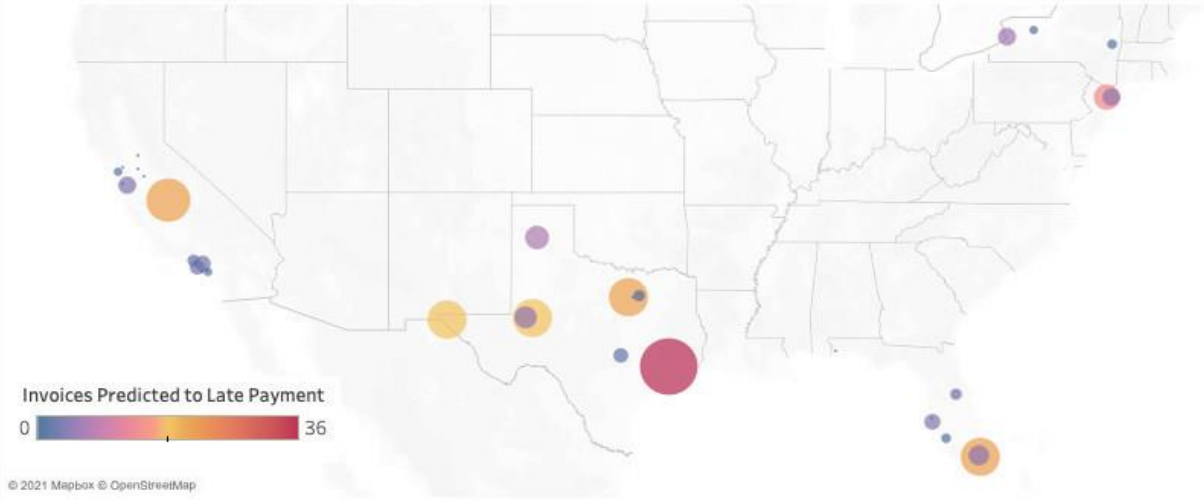
데이터 분석의 결과를 한눈에 볼 수 있도록 시각화



[출처] https://www.prnewswire.com/news-releases/tableau-business-science-brings-powerful-data-science-capabilities-to-business-people-301253805.html

$145,719	$20,900	2,437	13%	552
Invoice amount	Amount Impacted by Late Payment (Predicted)	Number of Invoices	% Late Payment Predicted	Number of Disputed Invoices

Predicted Delays Per City | Circle size shows amount impacted | Hover over to see details



Accounts With Late Payment Predicted

Sort By: Number of Predicted Late Invoices

Account Name	Delayed Payments (% total)	Invoices Predicted to Delayed Payment
Voomm	61%	17
Browsedriv	56%	15
Jabberbean	58%	15
Thoughtsto	39%	15
Vitz	48%	15
Browsebug	42%	13
Meemm	34%	12
Eare	33%	11
Twitterbea	48%	11
Centizu	59%	10
Jabbercube	20%	10
Innojam	33%	9
Centimia	23%	8
Jetpulse	33%	7
Oyope	26%	7
Realcube	9%	7
Skyble	32%	7
Skyndu	30%	7
Yabox	32%	7
Aibox	14%	6
Brainloung	24%	6
Izio	12%	6
Zazio	19%	6
Divanoodle	24%	5
Gabtune	19%	5
Leexo	17%	5
Myworks	11%	5

32%

Likelihood to be paid on time

2,437 selected rows

Top Predictors

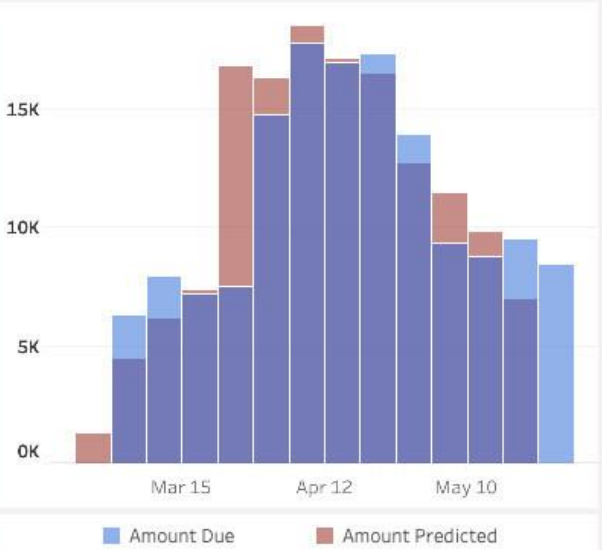
- 1.6 Disputed is No (960 rows)
- 0.6 Disputed is Yes (343 rows)
- 0.53 Payment term is 30 days (327 rows)

How To Improve This

- 16.96 Change Payment term to 60 days (1,002 rows)
- 9.79 Change Invoice type to Electronic (1,085 rows)

Powered by Einstein Discovery

Predicted Clashfow impacted by late payments



학습정리 Summary

1 빅데이터 프레임워크의 개념

- 빅데이터 수집, 저장, 처리, 분석, 시각화를 손쉽게 가능하도록 라이브러리 및 모듈로 기반 기술을 제공하는 소프트웨어

2 빅데이터 프레임워크의 기능

- 빅데이터의 수집, 저장, 처리, 분석 및 시각화를 위한 기능을 함
- 각 기능에 대한 여러 종류의 프레임워크가 존재함

참고문헌 Reference

- 빅데이터 컴퓨팅 기술, 2014, 박두순, 한빛미디어

서체 출처 : 나눔스퀘어-㈜네이버, G마켓 산스-㈜G마켓