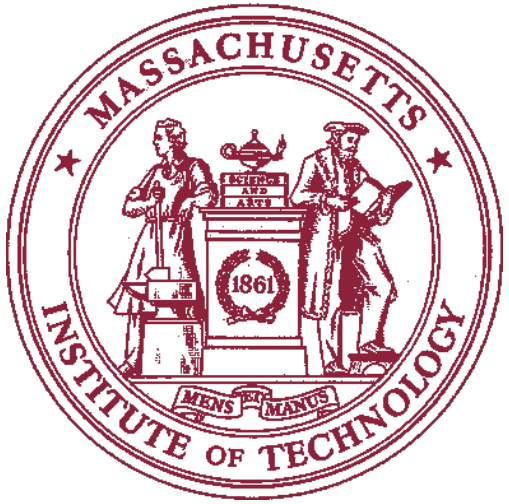
DataHub

A hosted platform for organizing, managing, sharing, collaborating, and processing data

Anant Bhardwaj

MIT CSAIL



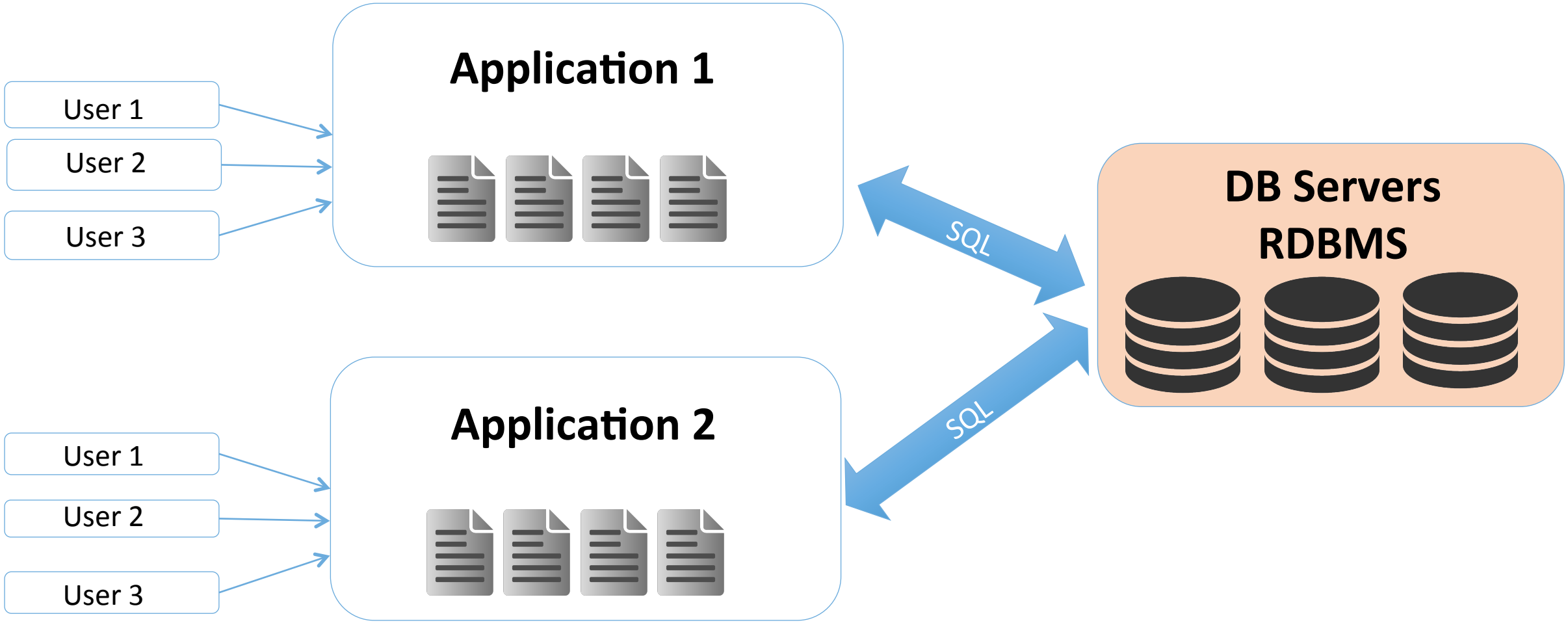
Anant Bhardwaj
Sam Madden
David Karger
Elizabeth Bruce
Eugene Wu

Aaron Elmore

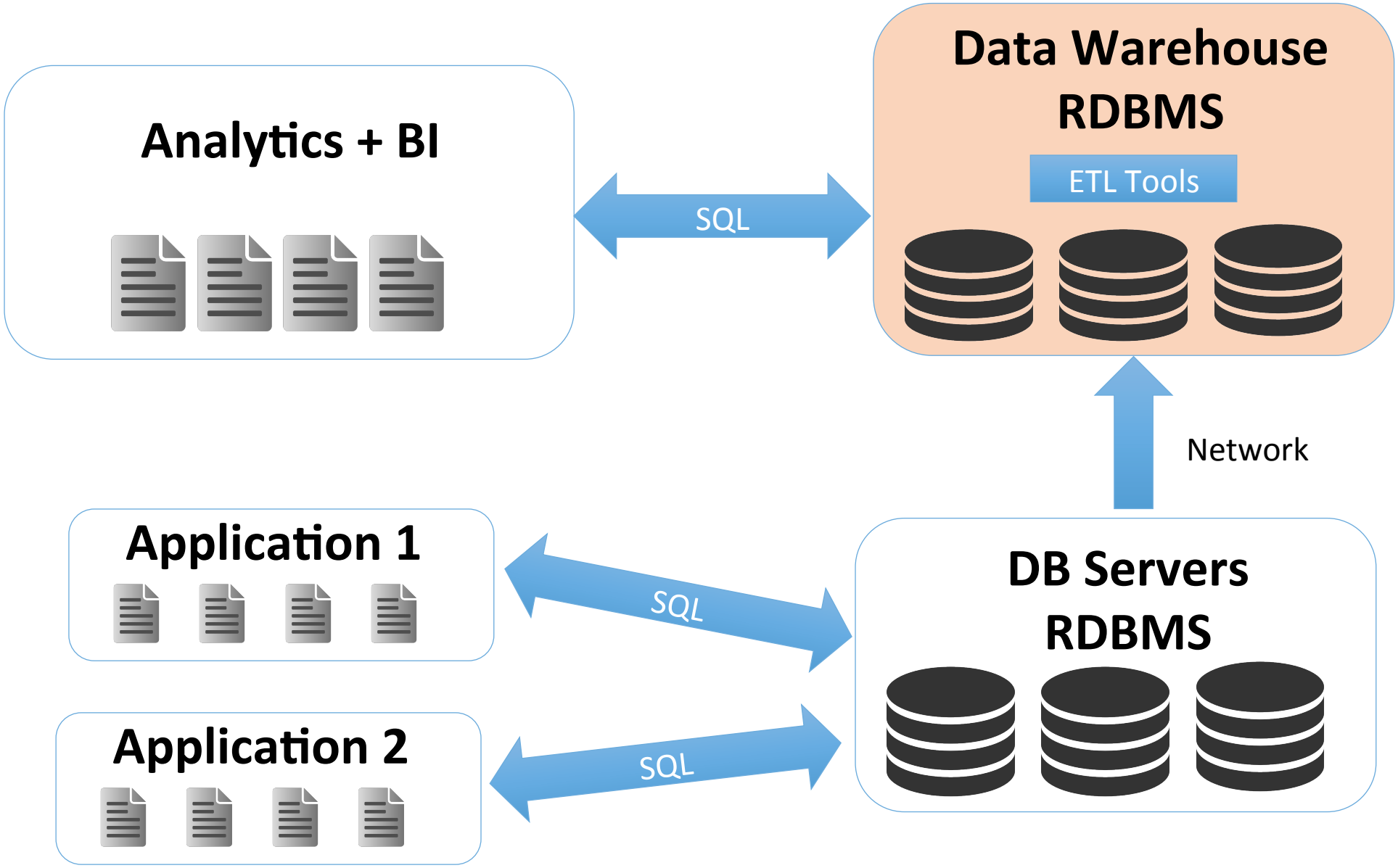
Aditya Parameswaran

Amol Deshpande

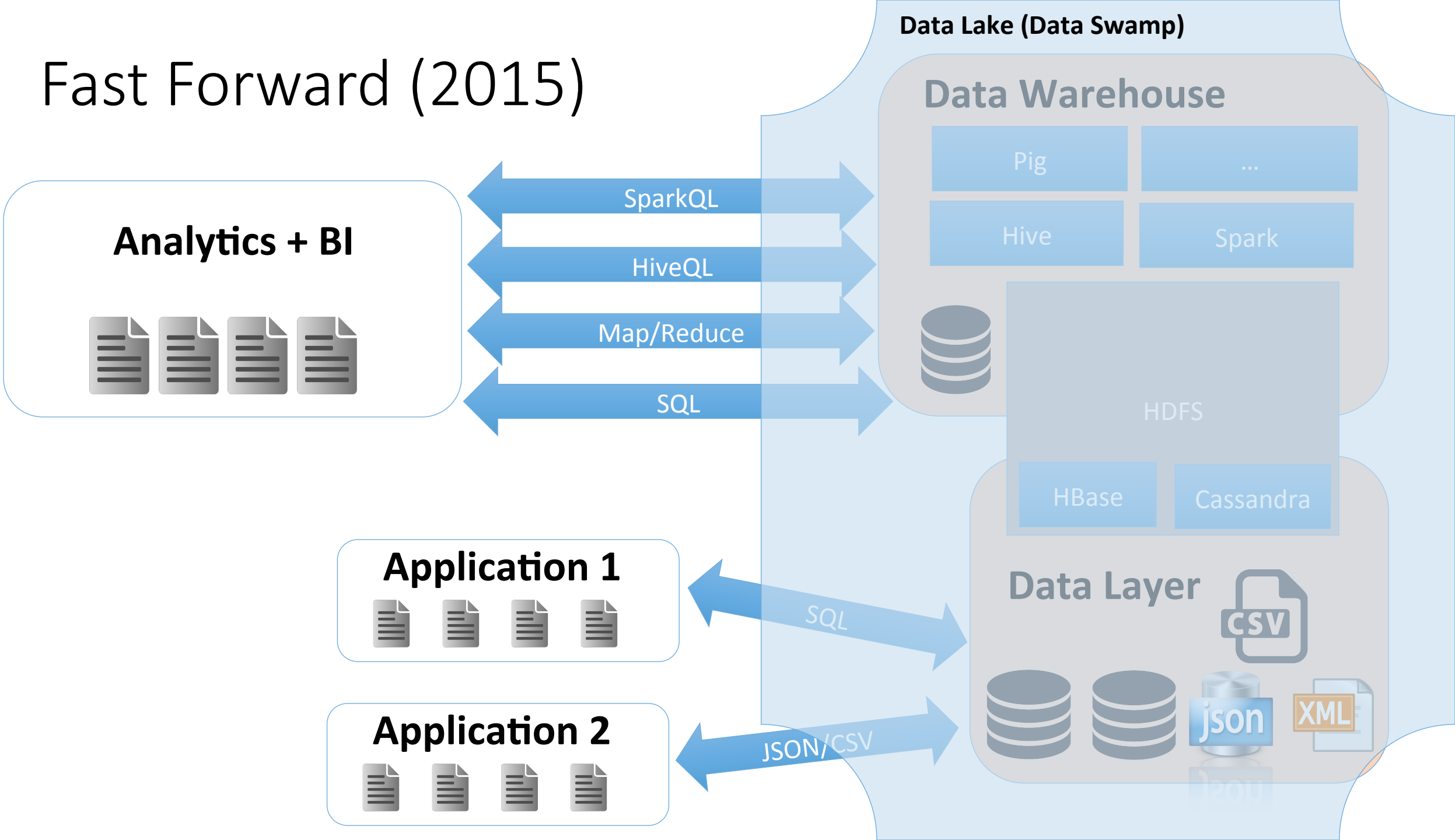
A bit of history: once upon a time



A bit of history: once upon a time



Fast Forward (2015)



An Information Provider's Wish List for a Next Generation Big Data End-to-End Information System

Mona M. Vernon
Thomson Reuters
22 Thomson Place
Boston, MA 02210, USA
mona.vernon@thomsonreuters.com

Brian Ulicny
Thomson Reuters
22 Thomson Place
Boston, MA 02210, USA
brian.ulicny@thomsonreuters.com

Dan Bennett
Thomson Reuters
610 Opperman Drive
Eagan, MN 55123, USA
dan.bennett@thomsonreuters.com

CIDR 2015

"The value of data is directly proportional to the degree to which it is accessible. As this data grows in size, in type, in dimensionality and in complexity, accessibility becomes of paramount importance.

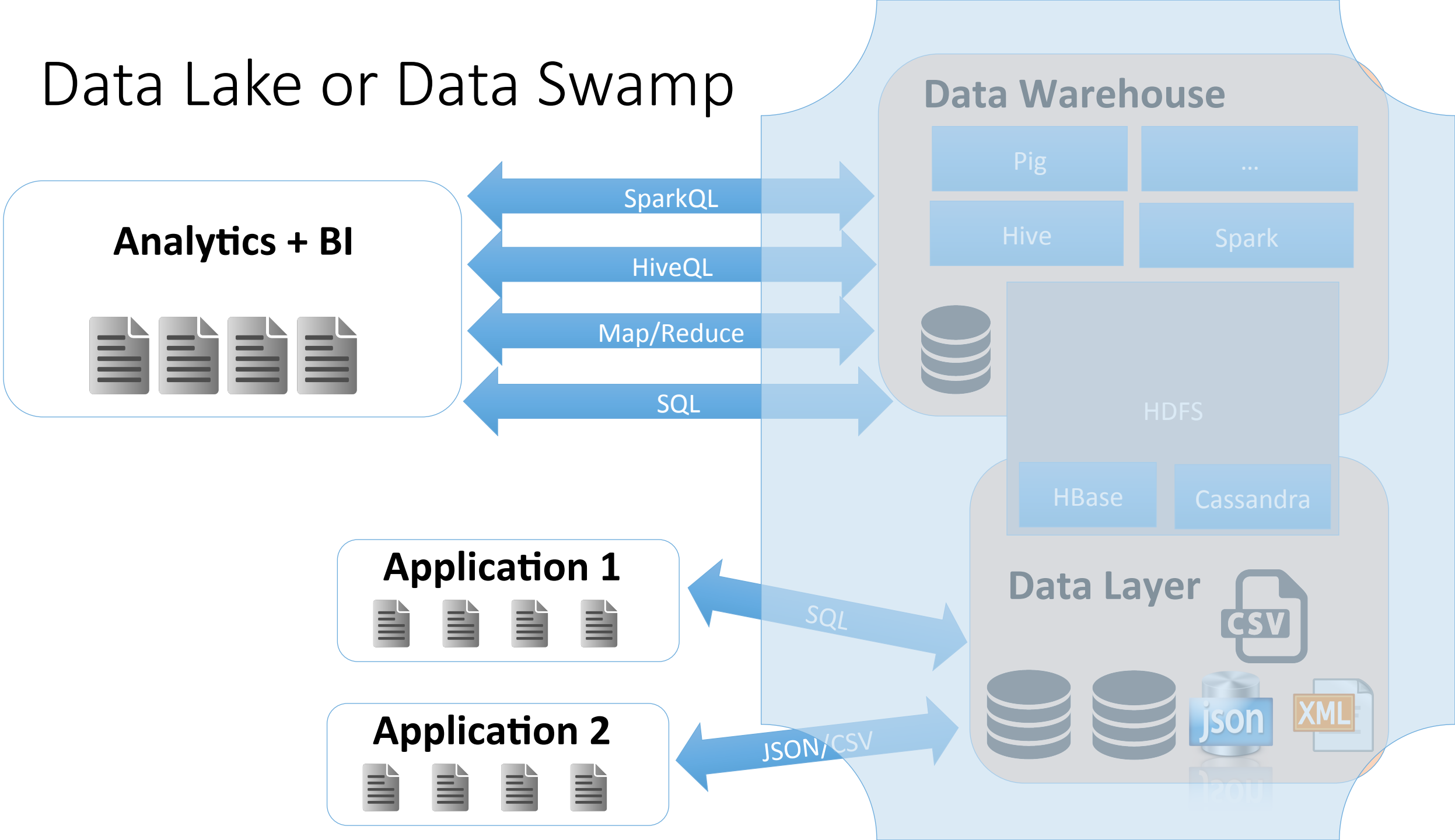
Accessibility means a very low barrier of entry: allowing product designers, innovators and non-technical users to explore and navigate the store without requiring them to be familiar with Big Data tools or in-depth understanding of data schemas and information models. Search and navigation support for identifying the data necessary to solve a particular analytical problem is lacking**."**

Accessibility means a very low barrier of entry

allowing
product designers, innovators and non-technical users to explore and navigate the store

without
requiring them to be familiar with Big Data tools, data schemas and information models

Data Lake or Data Swamp



Lessons

- End users don’t care about Data Models
  - They care about tasks
  - They will choose tools that can get their tasks done

- Supporting end-user tasks is very important
  - People love to use MS Excel, Google Sheets to whatever extent they can

Can we design Data Management System keeping end-users in mind?

Who are the end users

- Programmers
  - They chose data-management system that best fits their app (javascript developer is likely to choose JSON based data-model)
- Data Scientists
  - They care about extracting information from the data – data analysis
- Business Owners
  - They care about numbers/charts/graphs – data-driven decision making
- Data Administrators
  - They care about privacy, access control, compliance

A sample user class: Journalists

- Typical Problems
  - How to get the data files in from data.gov to excel
    - Ingest problem
  - How to query (especially by linking data across files)
    - Query problem (SQL is not the answer)
  - How to quickly make sense of it
    - a visualization that summarizes the data

Adam F. Hutton <adamfhutton@gmail.com>

to David, Anant

Hello David and Anant,

I have successfully uploaded several .csv files to DataHub. But I'm still having difficulty uploading .txt files.

I'm probably making another silly mistake, as I was with naming my .csv tables.

I click on "Create New Table" then "Extract From Text" but when I paste the .txt file into the data field, the page just goes gray. Any suggestions?

I'd still like to have a phone call today if possible. Does 1 p.m. still work?

Adam F. Hutton
"Where Media and Politics Intersect"
A curated news magazine covering the convergence
of journalism and American politics.
917.623.0224

...

Adam F. Hutton <adamfhutton@gmail.com>

10/29/14

to Anant, Eirik, David, Janos

Questions about individual_contributions_2014:

The Individual Contributions tables are a little more cryptic than some of the others because contributions are not listed according to candidate, only by committee ID number. I've listed the committee ID numbers I'm interested in below, but will use the names to pose questions.

Also there are no column headers, only numbers. In the questions I will say which columns I'm interested in. Column 1 is the committee ID number -- so all these questions will involve column1

These are also more complex questions because I'm looking for overlaps. And overlaps of a specific kind, which is evident in my first two questions.

- **Which** individuals (column7) gave $2,500 **OR** more (column14) to Steve Israel (C00358952) **AND** $5,000 **OR** more (column14) to the DCCC (C00000935)? **What** is the individual's profession (column11) and **who** do they work for (column12)? **When** did they donate (column13) and **how much** (column14)?
- **Which** individuals (column7) gave $10,000 OR more (column14) to the DCCC (C00000935) **AND** $10,000 OR more (column14) to the House Majority PAC (C00495028) **AND** any of the candidates listed in the key? **What** is the individual's profession (column11) and **who** do they work for (column12)? **When** did they donate (column13) and **how much** (column14)?
- **Which** individuals (column7) gave any amount to **TWO OR MORE** of the candidates listed in the key? **What** is the individual's profession (column11) and **who** do they work for (column12)? **When** did they donate (column13) and **how much** (column14)?

Key:
Committees
C00000935 -- Democratic Congressional Campaign Committee
C00495028 -- House Majority PAC

Candidates
C00358952 -- Steve Israel
C00500736 -- Steve Israel
C00551226 -- Alex Sink
C00551390 -- Alex Sink
C00544221 -- Nick Casey
C00547166 -- Jennifer Garrison
C00543611 -- Kevin Strouse
C00510461 -- Pete Aguilar
C00417550 -- Dan Maffei
C00567727 -- Emily Cain
C00512129 -- Ron Barber

On Thu, Feb 27, 2014 at 1:05 PM, Marton, Janos (MORELAND) <Janos.Marton@moreland.ny.gov> wrote:
Hi Anant,

We have some commissioners coming in tomorrow, and it would be great to show them what DataHub's potential is. Even if we don't have the BOE data uploaded by them, just showing any cross query might be worth showing them. For example, if I can show any *hits* between names in the client field of the JCOPE client database and OSC state contracts database, that would be of interest. Can you send me the SQL query that would prompt that, and I can try running it?

Thanks,

Janos

****

Janos Marton
Moreland Commission to Investigate Public Corruption
212-417-4288
917-848-6915 (cell)

Introducing DataHub

a unified data management and collaboration platform for making data-processing easy

DataHub Platform

- a flexible data store (files, relational databases, extendible to other backends) with sharing/collaboration capabilities, managed on behalf of different users/groups

- an app ecosystem that hosts apps for various data-processing activities
  - apps for ingestion, curation, integration, analytics, visualization, and machine learning
  - a new application can be written/published to the DataHub App Center using our thrift-SDK (20+ supported languages)
  - the DataHub users can use any of the apps from the App Center for processing their data as it fits their need

DataHub Platform

App Ecosystem

App Ingest/ETL

App Data Science

App Visualization

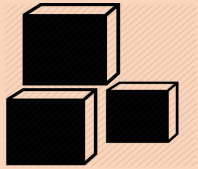
App Clustering

App Sentiment Analysis

App .....

DataHub Core

Users/Groups



Repository/ Catalogs

Sharing/Collaboration

Versioning

Data

Data

Data

Demo Time

Accessibility means a very low barrier of entry

allowing
product designers, innovators and non-technical users to
explore and navigate the store

without
requiring them to be familiar with bigdata tools, data schemas and information models

The real challenge today is not the data itself

but

an ecosystem for solving data-processing tasks

which end-users can use

DataHub: an ecosystem for end-users

- Powerful, easy to use collaboration
  - collaboration on data has never been so easy before
- Non Technical users can do stuff
  - it gives non-technical users the ability to find apps that best fits their data-processing needs
  - they can find apps for clustering, sentiment analysis, prediction, data cleaning, or anything from the app center
- Data Scientists can use tools/algorithms of their choice
  - data scientists can use R, Matlab, Julia, Java, anything that they are most comfortable with
  - multiple data scientists can collaborate seamlessly because an intermediate result computed in one language can be used in other language (DataHub enables true collaboration)
- Programmers can use DataHub in almost any language (20+ thrift supported languages)
  - DataHub can be used as a real backend in any web application, iPhone app, Android app, scripting languages (Python, PHP, Ruby, etc.), and system languages (Java, C++, Go, etc.)

Architectural Challenges

- A new architecture for storing and managing large numbers of diverse datasets
  - managed on behalf of different users/groups with sharing/collaboration capabilities.

- An infrastructure for hosting a large number of data-processing apps (the app ecosystem).

- Automating large parts of data science pipeline by letting users connect DataHub apps.

Data Science + Design Challenges

- Enabling easy ingest, manage, clean, analyze, and visualize large volumes of potentially unstructured and noisy data.

- Designing the user-experience (human-computer interaction aspect) to enable non-experts, such as social scientists and journalists, to effectively use the system and complete their tasks.

- Providing powerful capabilities for expert users to perform more complex reasoning tasks and analytics than what is possible today.