Big Data Analytics Platform

National Workshop on
Recent Trends in Big Data Analytics
February 27, 2016

All trademarks, service marks, trade names, product names and logos appearing on the documents are the property of their respective owners. Copyright (c) 2015 Computer Society of India. The document is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

16 Top Big Data Analytics Platforms

Copyright (c) 2016 Computer Society of India. Licensed under CC BY-NC-SA 4.0

Watson Data Analytics

Why do you need?

- **To answer business questions**
  - Use your own words to explore and visualize your data
- **To take confident action on insights**
  - Learn what's most likely to influence outcomes
- **To tell a compelling story**
- Build a dashboard or an infographic and let the stories come. Choose a template and add the information you want your audience to understand.
- **To analyze trusted data**
  - One can trust what you're using for your insights.

A smart data discovery service available on the cloud, it guides data exploration, automates predictive analytics and enables effortless dashboard and infographic creation.

How does it work?

Start with Data

- It can be a spreadsheet on your computer, relational data in a database, report data or data you've stored in a cloud storage service. You upload it or connect to it – and almost immediately – Watson Analytics provides you with interesting insights in the form of questions you can investigate.

Explore

- Do you want to get a better understanding of your data? Click Explore, choose a recommended question—or ask your own—and watch it come to life in visualizations.

Predict

- Do you want to determine the strongest influences on a specific outcome? Click Predict, select a target outcome, and see what is most likely to affect it or even make it happen.

Assemble

- Do you want to put different aspects of your data together in a dashboard or infographic? Click Assemble, get a template you can drag and drop data into and tell your story.

Copyright (c) 2016 Computer Society of India. Licensed under CC BY-NC-SA 4.0

Google Analytics

GA can show user behavior

- How do patrons find catalog page
  - Keywords
  - Referrers
- How do patrons search in the catalog?
  - Types of searches (keyword, author, title)
  - Scopes, limits
  - Number of search terms
- How do patrons use My Millennium features?
- How do patrons use the records they find?
  - Track outbound links to e-resources
  - Track outbound links to library webpages

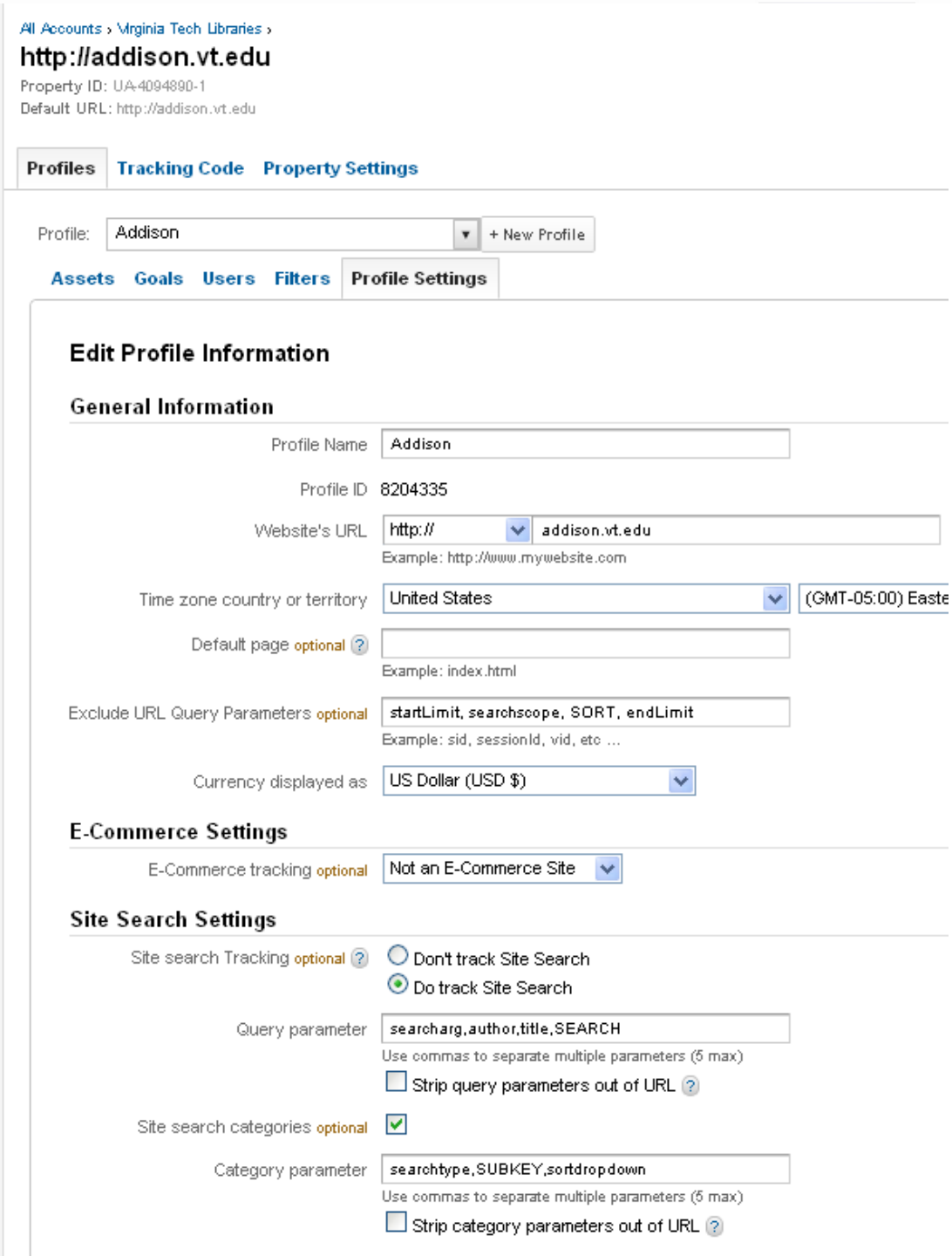
Data driven decisions

- What to include in help screens, prompt text on search pages, error pages
- What is our default search (keyword vs title)
- What do we include in catalog (if very low click trough's to an eBook package, do we rely on links from vendor site (database) or Web Scale Discovery Engine

What to look at in the future

- How does our implementation of Summon affect catalog usage?
  - Click throughs from Summon (referrers and record # searches)
  - Decline in number of searches?
- Does new emphasis on quality searches in instruction classes change number of search terms?
  - Decrease in single word searches and over 10 word searches?

Copyright (c) 2016 Computer Society of India. Licensed under CC BY-NC-SA 4.0

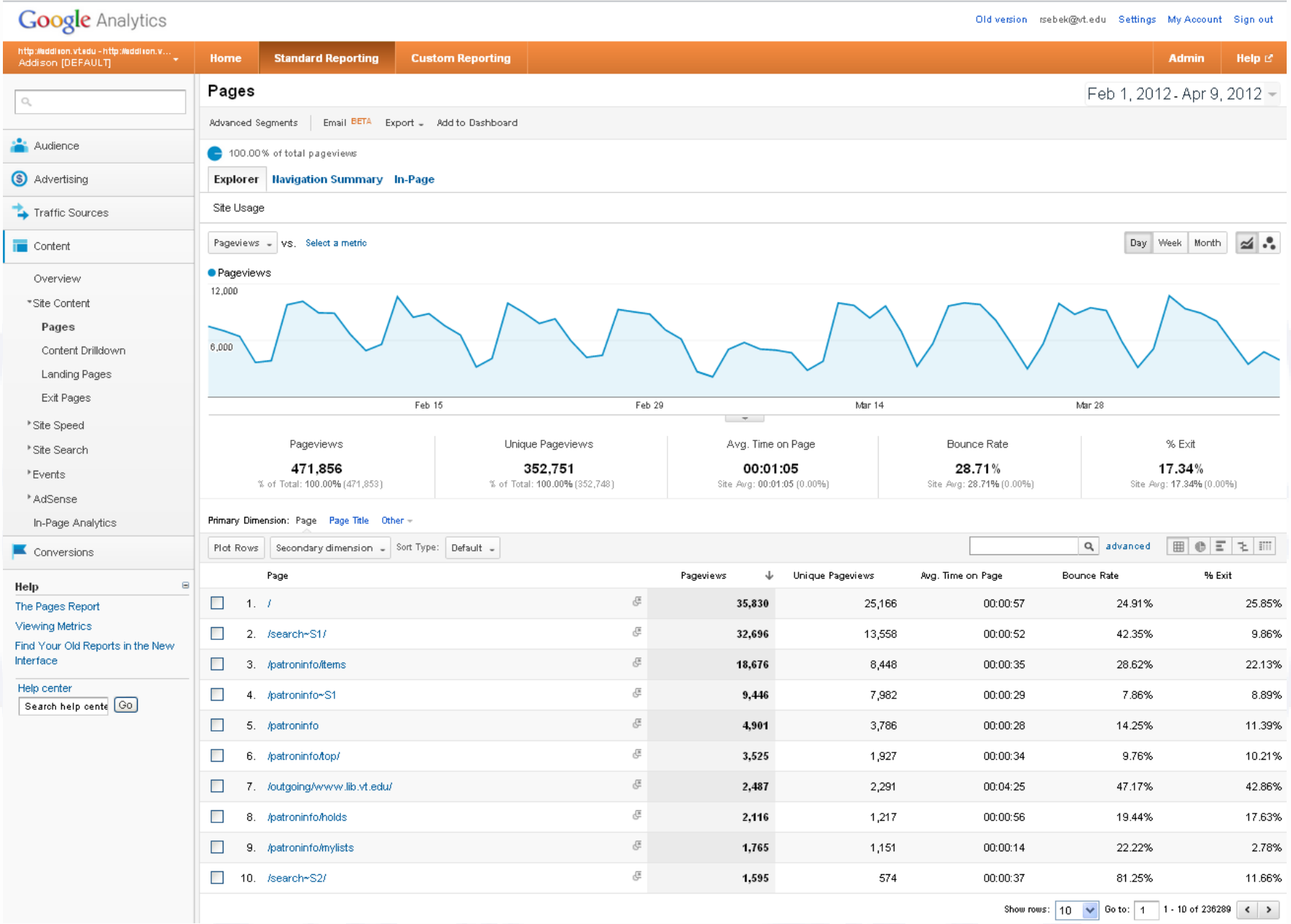


Create a GA profile

Configure Site Search Settings

Keep searcharg, author, title, SEARCH

Exclude startLimit, SORT, endLimit, (and maybe searchscope)



Top Content is the most commonly viewed report, so let's use it to define the labels.

Google | Analytics

My Account | Help | Contact us

Analytics Home | View Reports www.googlestore.com My Analytics Accounts Google Store

Reports: www.googlestore.com

Dashboards

- View: Executive
  - Executive Overview
  - E-commerce Summary
  - Conversion Summary
  - Marketing Summary
  - Content Summary
  - Site Overlay

All Reports

- Marketing Optimisation
- Content Optimisation
- E-Commerce Analysis

Date Range

View By: Default

2005

Jan	Feb	Mar	Apr	May	Jun
Jul	Aug	Sep	Oct	Nov	Dec

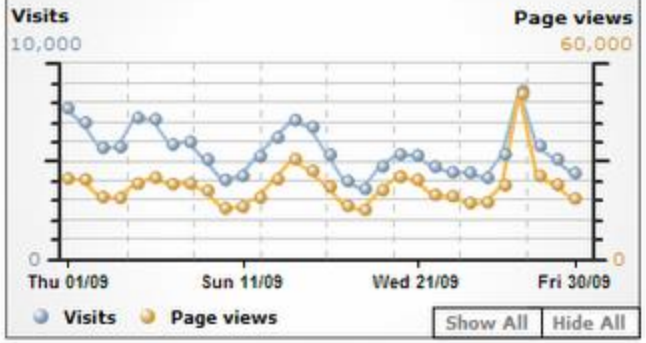
	S	M	T	W	T	F	S
→	28	29	30	31	1	2	3
→	4	5	6	7	8	9	10
→	11	12	13	14	15	16	17
→	18	19	20	21	22	23	24
→	25	26	27	28	29	30	1

Prev << Month >> Next

Executive Overview Export

www.googlestore.com | 01/09/2005 - 30/09/2005

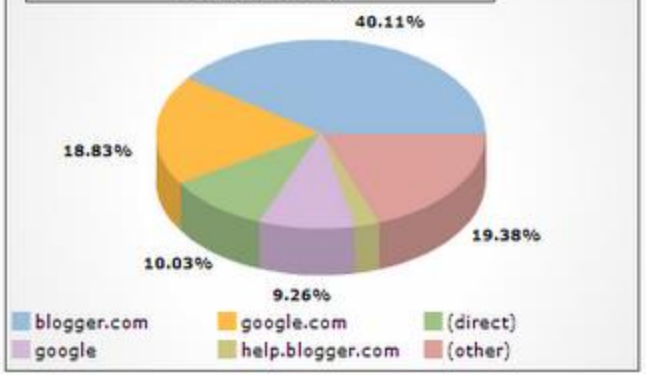
Visits and Page views



Visits by New and Returning

Geo Map Overlay

Visits by Source



Help Information

Visitor Summary The four graphics in this report provide a quick snapshot of visits to your site. The following are shown:

Copyright (c) 2016 Computer Society of India. Licensed under CC BY-NC-SA 4.0

Google Cloud DataFlow

- Google Cloud Dataflow is a simple, flexible, and powerful system you can use to perform data processing tasks of any size

Cloud Dataflow consists of two major components:

- **A set of SDKs that you use to define data processing jobs.** The Dataflow SDKs feature a unique programming model that simplifies the mechanics of large-scale cloud data processing. You can define your data processing jobs by writing programs using the Dataflow SDKs.
- **A Google Cloud Platform managed service.** The Dataflow service ties together and fully manages several different Google Cloud Platform technologies, such as Google Compute Engine, Google Cloud Storage, and BigQuery to execute data processing jobs on Google Cloud Platform resources.

Dataflow SDK Concepts

Key concepts in the Dataflow SDKs include:

- **Simple data representation.** Dataflow SDKs use a specialized collection class, called PCollection, to represent your pipeline data. This class can represent data sets of virtually unlimited size, including bounded and unbounded data collections.
- **Powerful data transforms.** Dataflow SDKs provide several core data transforms that you can apply to your data. These transforms, called PTransforms, are generic frameworks that apply functions that you provide across an entire data set, using the features of the Dataflow service to execute each transform in the most efficient way.
- **I/O APIs for a variety of data formats.** Dataflow SDKs provide APIs that let your pipeline read and write data to and from a variety of formats and storage technologies. Your pipeline can read text files, Avro files, BigQuery tables, and more.

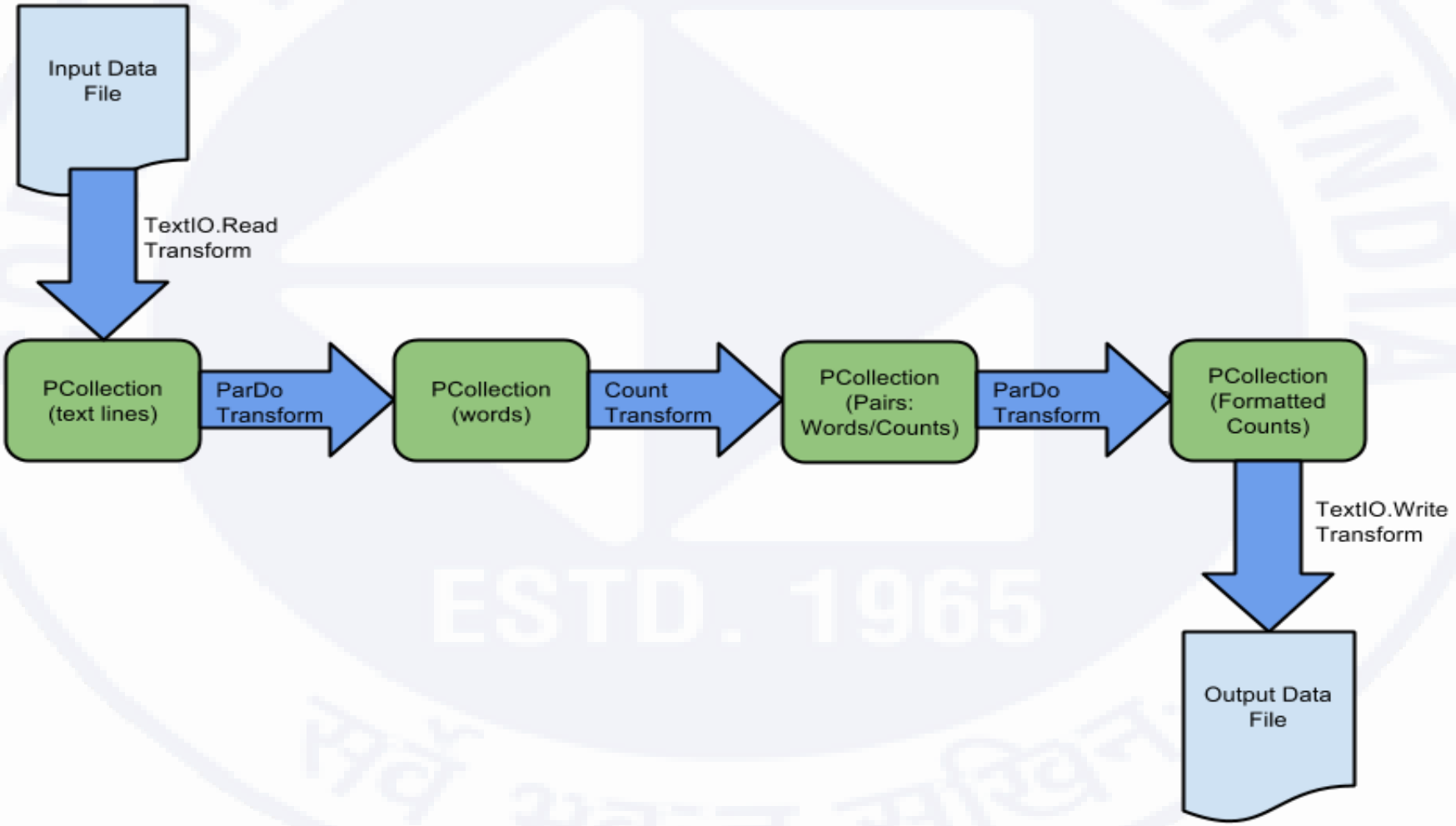
Benefits of an open source Dataflow ecosystem

- **Increased visibility**. The Dataflow SDK source code provides insight on how Dataflow programs interact with the managed Dataflow service that Google provides.
- **Support for third-party pipeline runners**. Releasing the Dataflow SDKs as open source makes it easier for others to provide their own services to run pipelines defined using the Dataflow SDKs. These runners might target different runtime environments or enable Dataflow pipelines to be run on premises, using non-Google Cloud Platform services.
- **Transform Modeling**. The pre-written transforms in the Dataflow SDKs can provide models or design patterns that the open source community can use to contribute their own transforms to the ecosystem.

Data Flow Service Features

- **Dynamic Optimization:** The Dataflow service provides dynamic optimization of Cloud Platform resources to execute your data processing jobs. When you build a dataflow, the Dataflow service constructs a directed graph of your job and optimizes the graph for the most efficient execution.
- **Resource Management:** The Dataflow service fully manages Cloud Platform technologies to run your job. This includes spinning up and tearing down Compute Engine resources, collecting logs, and communicating with Cloud Storage technologies.
- **Job Monitoring:** The Dataflow service includes a monitoring interface built into the Google Cloud Platform Console. The Dataflow monitoring interface shows the different stages of your data processing pipeline and lets you see how data moves through those stages as the job progresses.
- **Native I/O Adapters for Cloud Storage Technologies:** The Dataflow Service has built-in support for getting data from, and writing data to, Cloud Platform storage systems such as Cloud Storage and BigQuery. This makes it easy to build a data processing pipeline to work with your data in Cloud Platform.

Pipeline Transformation



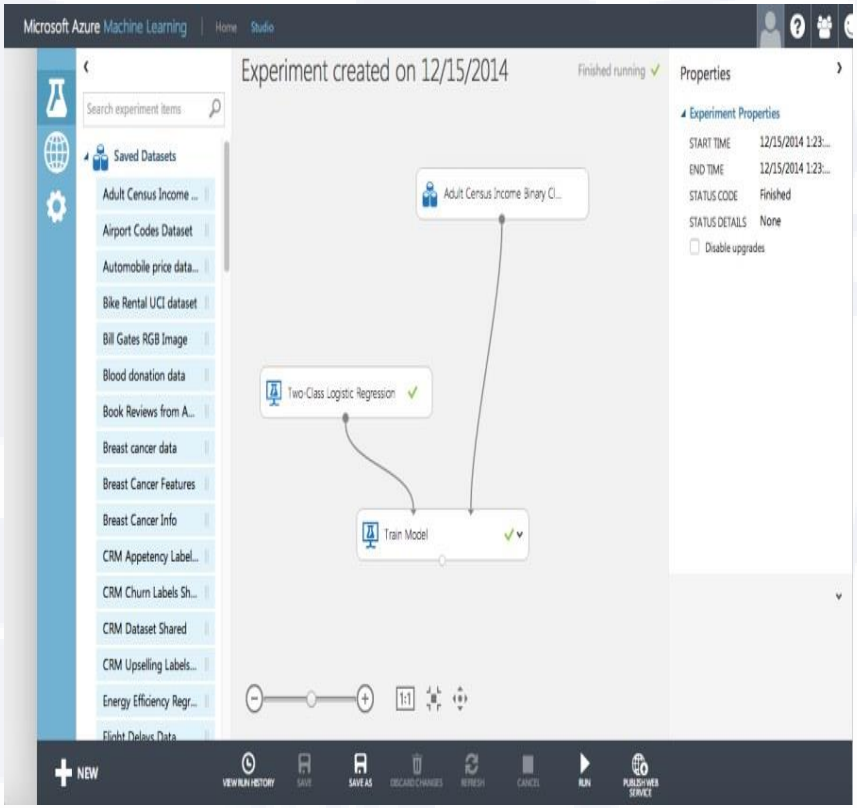
Copyright (c) 2016 Computer Society of India. Licensed under CC BY-NC-SA 4.0

Microsoft Azure

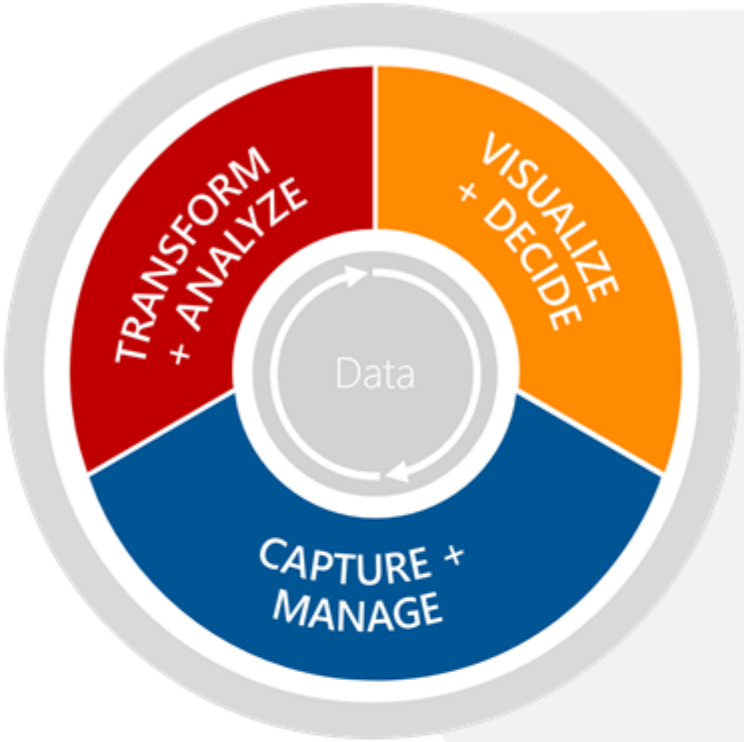
Machine Learning Tool

- Microsoft's Azure Machine Learning software is a tool that automates some tasks in the machine learning pipeline, assuming familiarity with basic data science techniques. It presents a user-friendly, graphical drag and drop interface to route data through various preprocessing steps and ultimately into a machine learning algorithm.

Sample Canvas



Copyright (c) 2016 Computer Society of India. Licensed under CC BY-NC-SA 4.0



Microsoft SQL Server
Microsoft Azure
Office

The Microsoft data platform

Visualize + decide

- Applications
- Reports
- Dashboards

Transform + analyze

- Orchestration
- Information management
- Complex event processing

Capture + manage

- Relational
- Non-relational
- NoSQL

Copyright (c) 2016 Computer Society of India. Licensed under CC BY-NC-SA 4.0

Watson Analytics Vs. Microsoft Azure

- *IBM Watson Analytics prototype seeks to abstract away data science, taking ordinary natural language queries and answering them based on the content of uploaded datasets. Microsoft Azure Machine Learning goes the opposite route, streamlining existing data mining methodology for fast results and integration with MS's other cloud services.*
- Fundamentally different use cases they address
- Watson Analytics and Microsoft's Azure Machine Learning present very different products, albeit with some overlapping features.
- While IBM seeks to make it possible for anyone at all to interrogate data.
- Azure on the other hand sets the more modest goal of wrapping a user-friendly interface around machine learning tasks, and integrating machine learning into existing business workflows.

Copyright (c) 2016 Computer Society of India. Licensed under CC BY-NC-SA 4.0

Gartner 2016 Magic Quadrant for Advanced Analytics Platforms

- The 2016 report evaluated 16 analytics and data science firms over 10 criteria and placed them in 4 quadrants, based on completeness of vision and ability to execute.
- **Leaders (5):** SAS, IBM, KNIME, RapidMiner, Dell
- **Challengers (2):** SAP, Angoss
- **Visionaries (4):** Microsoft, Alteryx, Alpine Data Labs, Predixion Software
- **Niche Players (5):** FICO, Lavastorm, Megaputer, Prognoz, Accenture

Copyright (c) 2016 Computer Society of India. Licensed under CC BY-NC-SA 4.0