INTERACTIVE ANALYTICS DASHBOARD USING SUPERSET

PROJECT MENTOR

ZAMA

PROJECT MEMBER

MITHIL VS

USN:-1MV15CS060

DEPT OF CSE, SIR M VISVESVARAYYA INSTITUTE OF TECHNOLOGY

ACKNOWLEDGMENT

I give my high, respectful gratitude to my Mentor **Mr. BadiyuZama Mohammed**, who has been my source of inspiration. He has been especially enthusiastic in giving his opinions and critical reviews. I have learnt a lot throughout this internship with many challenges yet valuable experience to complete this task. I will remember his contribution forever.

I sincerely thank **Mr. Harish Daga**, who have been the constant driving force behind the completion of project. I also thank **Mr. Prasad Kothapalli** for his constant help and support throughout.

I also take this opportunity to thank the staffs who have helped us a lot in providing the guidance and any kind of help whenever needed. My thanks and appreciations also go to my friends who have willingly helped us with their abilities.

-Mithil VS

TABLE OF CONTENTS

1.0 DATA GENERATORS

We can produce three types of data

- Structured data
- Unstructured data
- Semi structured data

1.1 Structured data

Structured data usually resides in relational databases (RDBMS). Fields store length-delineated data phone numbers, Social Security numbers, or ZIP codes. Even text strings of variable length like names are contained in records, making it a simple matter to search. Data may be human- or machine-generated as long as the data is created within an RDBMS structure. This format is eminently searchable both with human generated queries and via algorithms using type of data and field names, such as alphabetical or numeric, currency or date Common relational database applications with structured data include airline reservation systems, inventory control, sales transactions, and ATM activity. Structured Query Language (SQL) enables queries on this type of structured data within relational databases. Some relational databases do store or point to unstructured data such as customer relationship management (CRM) applications. The integration can be awkward at best since memo fields do not loan themselves to traditional database queries. Still, most of the CRM data is structured.

1.2 Unstructured Data

Most IT workers are used to structured data. It is written in a format that's easy for machines to understand, though it baffles most people unless they're programmers. Structured data is easily searchable by basic algorithms. Examples include spreadsheets and data from machine sensors.

Unstructured data is more like human language. It doesn't fit nicely into relational databases like SQL, and searching it based on the old algorithms ranges from difficult to completely impossible. Examples include emails, text documents (Word docs, PDFs, etc.), social media posts, videos, audio files, and images. There is a lot of important, useful information locked up in all that unstructured data. The information in emails and social media, for example, holds important insight that can be used for operational intelligence, marketing intelligence, and more. This kind of information can tell businesses things

beyond a customer survey, such as what the public has to say about your latest products or changes in store hours. It also holds information on the production process, various ongoing projects, plans for the future, and much more. Pictures from your last R&D project, for instance, might be helpful in generating better ideas for creative endeavours down the road.

There are number of ways to begin assembling a database of unstructured data and processing it to yield this valuable insight. Many companies have migrated to an object-oriented database such as NoSQL. Some have also undertaken open source big data analysis initiatives, like Hadoop.

1.3 Apache Kafka

This data is given to **Apache Kafka** which is an open-source stream-processing software platform developed by the Apache Software Foundation, written in Scala and Java. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds. Its storage layer is essentially a "massively scalable pub/sub message queue architected as a distributed transaction log, making it highly valuable for enterprise infrastructures to process streaming data. Additionally, Kafka connects to external systems (for data import/export) via Kafka Connect and provides Kafka Streams, a Java stream processing library.

2.0 DATA ANALYZERS

2.1 DRUID

Druid is a column-oriented, open-source, distributed data store written in Java. Druid is designed to quickly ingest massive quantities of event data, and provide low-latency queries on top of the data. The name Druid comes from the shapeshifting Druid class in many role-playing games, to reflect the fact that the architecture of the system can shift to solve different types of data problem. Druid is commonly used in business intelligence/OLAP applications to analyse high volumes of real-time and historical data.

2.2 OLAP

Online analytical processing, or **OLAP** is an approach to answering multi-dimensional analytical (MDA) queries swiftly in computing. OLAP is part of the broader category of business intelligence, which also encompasses relational databases, report writing and data mining. Typical applications of OLAP include marketing,management ,budgeting and forecasting, financial

reporting and similar areas, with new applications coming up, such as agriculture. The term *OLAP* was created as a slight modification of the traditional database term online transaction processing (OLTP).

OLAP tools enable users to analyse multidimensional data interactively from multiple perspectives. OLAP consists of three basic analytical operations: consolidation (roll-up), drill-down, and slicing and dicing. Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions. For example, all sales offices are rolled up to the sales department or sales division to anticipate sales trends. By contrast, the drill-down is a technique that allows users to navigate through the details. For instance, users can view the sales by individual products that make up a region's sales. Slicing and dicing is a feature whereby users can take out (slicing) a specific set of data of the OLAP cube and view (dicing) the slices from different viewpoints. These viewpoints are sometimes called dimensions (such as looking at the same sales by salesperson or by date or by customer or by product or by region, etc.)

Databases configured for OLAP use a multidimensional data model, allowing for complex analytical and ad hoc queries with a rapid execution time. They borrow aspects of navigational databases, hierarchical databases and relational databases.

OLAP is typically contrasted to OLTP (online transaction processing), which is generally characterized by much less complex queries, in a larger volume, to process transactions rather than for the purpose of business intelligence or reporting. Whereas OLAP systems are mostly optimized for read, OLTP must process all kinds of queries (read, insert, update).

TYPES OF OLAPS

Multidimensional OLAP

Relational OLAP (ROLAP)

2.2.1 Open-source Servers of OLAP

- Mondrian OLAP server is an open-source OLAP server written in Java. It supports the MDX query language, the XML for Analysis and the olap4j interface specifications.
- Druid (open-source data store) is a popular open-source distributed data store for OLAP queries that is used at scale in production by various organizations.
- Apache Kylin is a distributed data store for OLAP queries originally developed by eBay.

- Cubes (OLAP server) is another light-weight open-source toolkit implementation of OLAP functionality in the Python programming language with built-in ROLAP.
- LinkedIn Pinot is used at LinkedIn to deliver scalable real-time analytics with low latency. It can ingest data from offline data sources (such as Hadoop and flat files) as well as online sources (such as Kafka). Pinot is designed to scale horizontally.

2.3 NOSQL

NoSQL (originally referring to "non SQL" or "non-relational") database provides a mechanism for storage and retrieval of data that is modelled in means other than the tabular relations used in relational databases. Such databases have existed since the late 1960s, but did not obtain the "NoSQL" moniker until a surge of popularity in the early twenty-first century, triggered by the needs of Web 2.0 companies. NoSQL databases are increasingly used in big data and real-time web applications. NoSQL systems are also sometimes called "Not only SQL" to emphasize that they may support SQL-like query languages.

Motivations for this approach include: simplicity of design, simpler "horizontal" scaling to clusters of machines (which is a problem for relational databases), and finer control over availability. The data structures used by NoSQL databases (e.g. key-value, wide column, graph, or document) are different from those used by default in relational databases, making some operations faster in NoSQL. The particular suitability of a given NoSQL database depends on the problem it must solve. Sometimes the data structures used by NoSQL databases are also viewed as "more flexible" than relational database tables.

2.3.1 TYPES OF NOSQL

There have been various approaches to classify NoSQL databases, each with different categories and subcategories, some of which overlap. What follows is a basic classification by data model, with examples:

- **Column**: Cassandra, Druid, HBase, Vertica.
- **Document**: ApacheCouchDB, ArangoDB, BaseX, Clusterpoint, Couchbase, C osmos DB, IBM Domino, Mark Logic, MongoDB, Orient DB, Rethink DB
- **Key-value**: Aerospike, ArangoDB, Couchbase, Dynamo,FairCom c-treeACE, FoundationDB, InfinityDB, MemcacheDB, MUMPS, Oracle NoSQL Database, Orient DB, Redid, Sci DB, SDBM/Flat File DB, Zoo Keeper
- **Graph**: Allegro Graph, Arango DB, Infinite Graph, Apache Graph, Mark Logic, Neo4J, Orient DB, Virtuoso

2.4 Time Series Database

Time series database (TSDB) is a software system that is optimized for handling time series data, arrays of numbers indexed by time (a datetime or a datetime range). In some fields these *time series* are called profiles, curves, or traces. A time series of stock prices might be called a price curve. A time series of energy consumption might be called a load profile. A log of temperature values over time might be called a temperature trace.

Despite the disparate names, many of the same mathematical operations, queries, or database transactions are useful for analysing all of them. The implementation of a database that can correctly, reliably, and efficiently implement these operations must be specialized for time-series data.

TSDBs are databases that are optimized for time series data. Software with complex logic or business rules and high transaction volume for time series data may not be practical with traditional relational database management systems. Flat file databases are not a viable option either, if the data and transaction volume reaches a maximum threshold determined by the capacity of individual servers (processing power and storage capacity). Queries for historical data, replete with time ranges and roll ups and arbitrary time zone conversions are difficult in a relational database

A TSDB allows users to create, enumerate, update and destroy various time series and organize them in some fashion. These series may be organized hierarchically and optionally have companion metadata available with them. The server often supports a number of basic calculations that work on a series as a whole, such as multiplying, adding, or otherwise combining various time series into a new time series. They can also filter on arbitrary patterns defined by the day of the week, low value filters, high value filters, or even have the values of one series filter another.

2.4.1 List of time series databases

The following notable database systems provide support for time series data:

Name	License	Language
extreme DB	Commercial	SQL, Python, C / C++, Java, and C#

Name	License	Language
Graphite	Apache 2	Python
Riak-TS	Apache License 2.0	Erlang
RRDtool	GPLv2	C

2.5 Apache Hadoop

It is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model. Originally designed for computer clusters built from commodity hardware—still the common use—it has also found use on clusters of higher-end hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

The base Apache Hadoop framework is composed of the following modules:

- Hadoop Common – contains libraries and utilities needed by other Hadoop modules;
- Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;

- Hadoop YARN – introduced in 2012 is a platform responsible for managing computing resources in clusters and using them for scheduling users' applications.
- Hadoop MapReduce – an implementation of the MapReduce programming model for large-scale data processing.

3.0 APACHE SUPERSET

Apache Superset (incubating) is a modern, enterprise-ready business intelligence web application

3.1 Features

- A rich set of data visualizations
- An easy-to-use interface for exploring and visualizing data

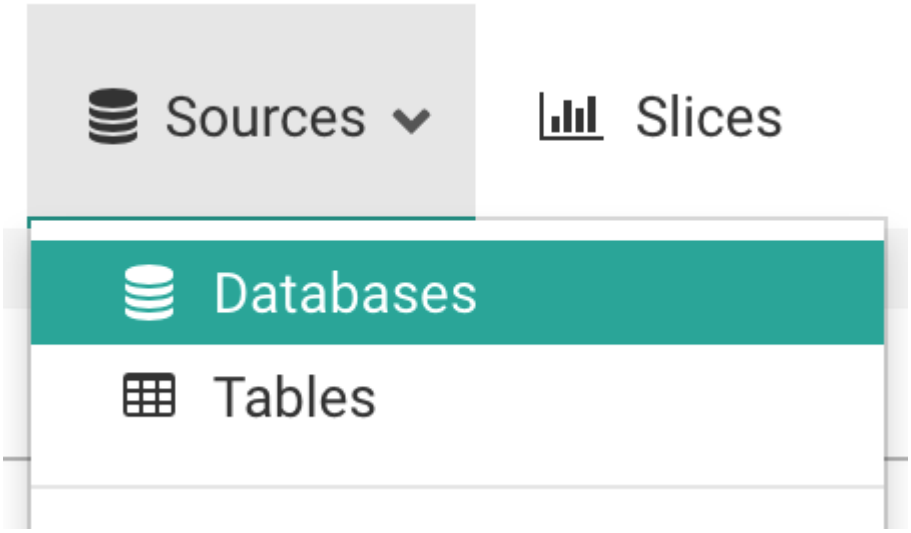
3.1 Create and share dashboards

- Enterprise-ready authentication with integration with major authentication providers (database, OpenID, LDAP, OAuth & REMOTE_USER through Flask App Builder)
- An extensible, high-granularity security/permission model allowing intricate rules on who can access individual features and the dataset
- A simple semantic layer, allowing users to control how data sources are displayed in the UI by defining which fields should show up in which drop-down and which aggregation and function metrics are made available to the user
- Integration with most SQL-speaking RDBMS through SQLAlchemy
- Deep integration with Druid.io

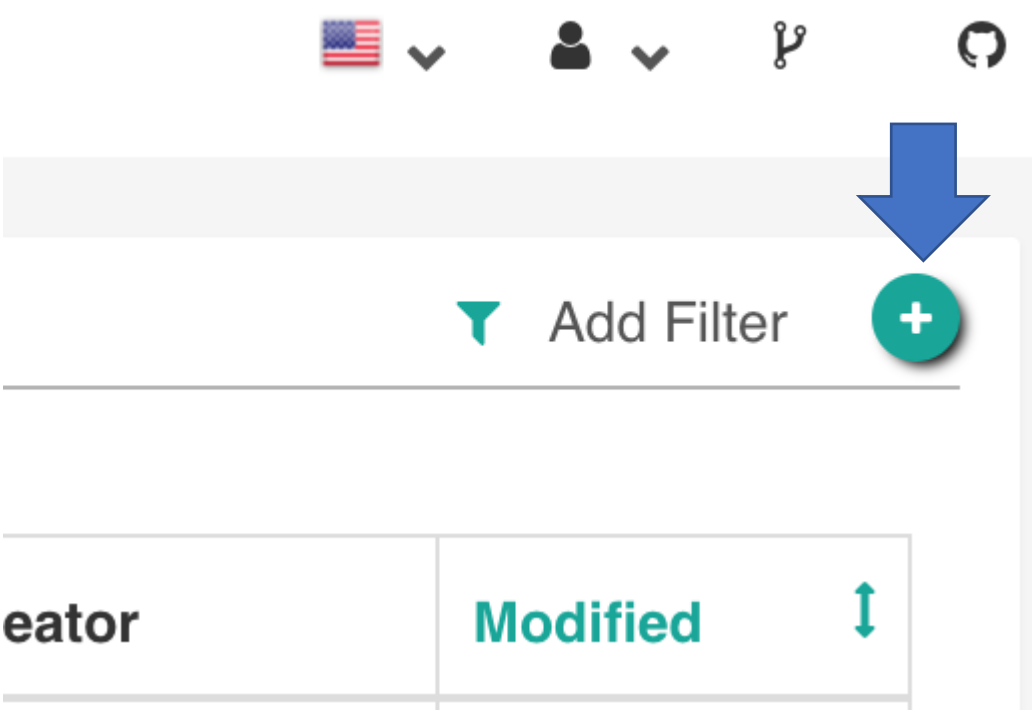
3.2 To Create a Database using Superset

We assume you already have a database configured and can connect to it from the instance on which you're running Superset.

Under the **Sources** menu, select the *Databases* option:



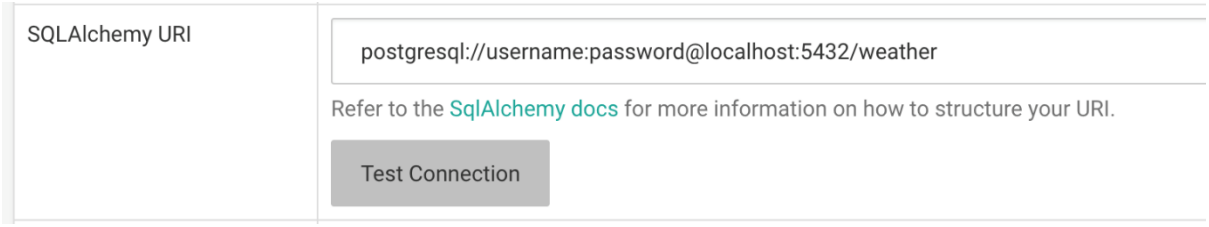
On the resulting page, click on the green plus sign, near the top right:



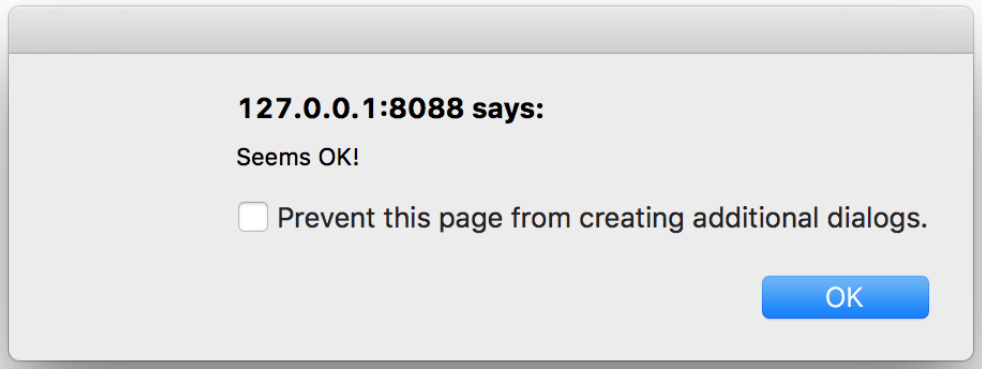
You can configure a number of advanced options on this page, but for this walkthrough, you'll only need to do **two things**:

1. Name your database connection:

2. Provide the SQLAlchemy Connection URI and test the connection:



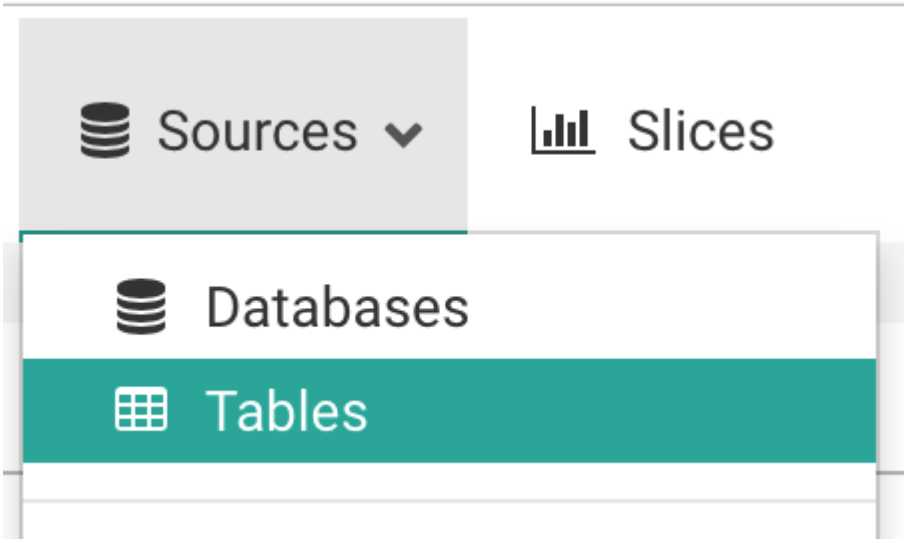
Click the **Test Connection** button to confirm things work end to end.



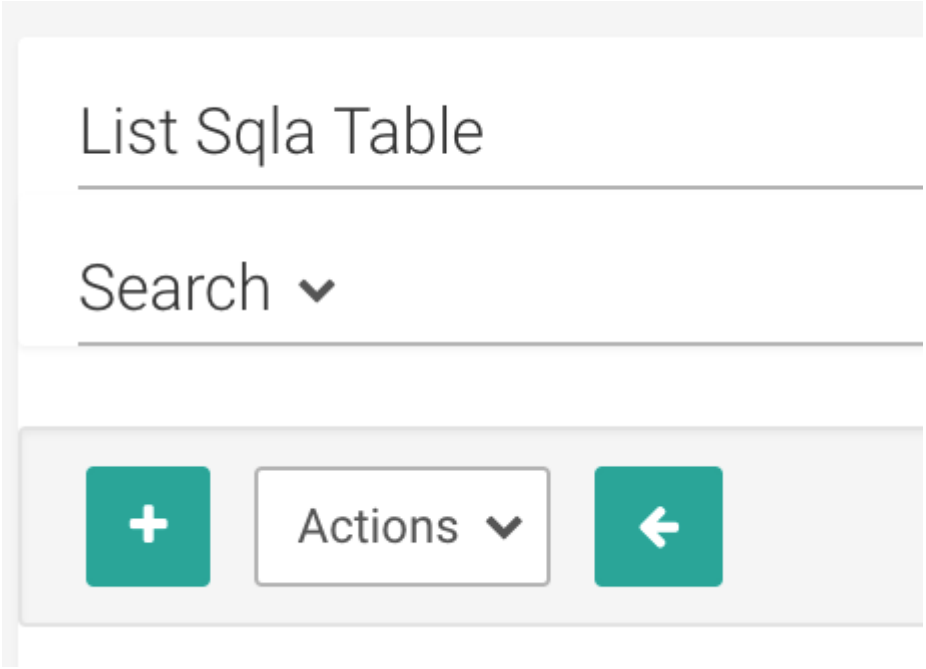
save the configuration by clicking the **Save** button at the bottom of the page.

3.3 Adding a new table

Under the **Sources** menu, select the *Tables* option:



On the resulting page, click on the green plus sign, near the top left:



You only need a few pieces of information to add a new table to Superset:

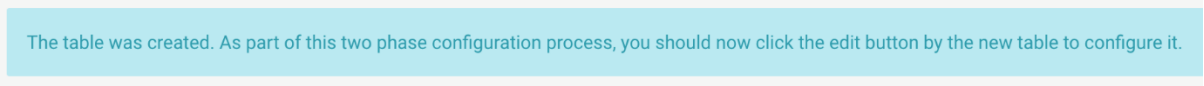
- The name of the table

The target database from the **Database** drop-down menu

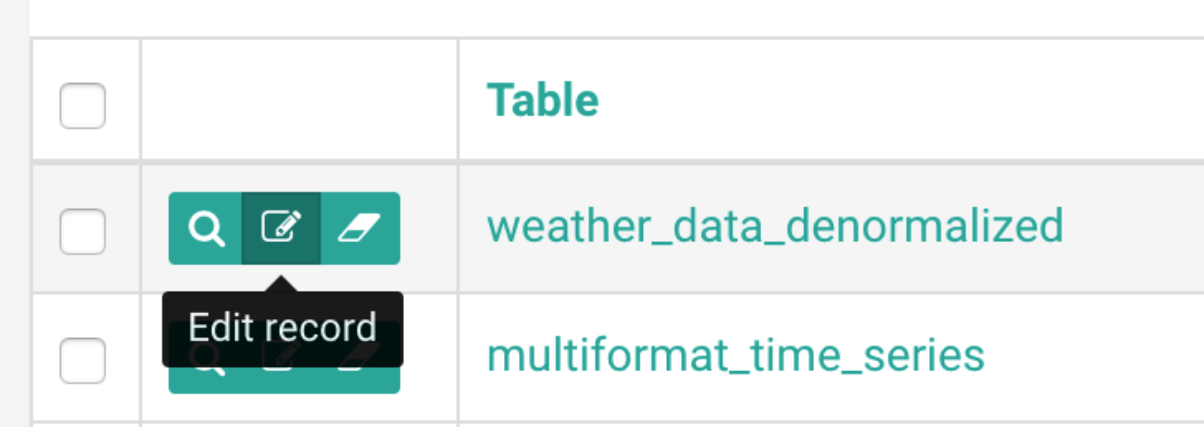
- Optionally, the database schema.

Click on the **Save** button to save the configuration:

When redirected back to the list of tables, you should see a message indicating that your table was created:



Click on the edit button next to the table you've created:



On the resulting page, click on the **List Table Column** tab. We'll run through these options to describe their purpose:

- If you want users to group metrics by a specific field, mark it as **Groupable**.
- If you need to filter on a specific field, mark it as **Filterable**.
- Is this field something you'd like to get the distinct count you can use **Count Distinct** box.
- Is this a metric you want to sum, or get basic sum The **Sum**, **Min**, and **Max** columns .
- The **is temporal** field should be checked for any date or time field.
    - As with the configurations above, click the **Save** button to save these settings.

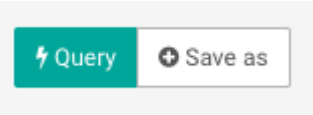
3.4 Exploring your data

To start exploring your data, simply click on the table name you just created in the list of available tables. By default, you'll be presented with a Table View. Let's group this by the *description* field to get the count of records by the type of record by adding it to the *Group by* section and run the query

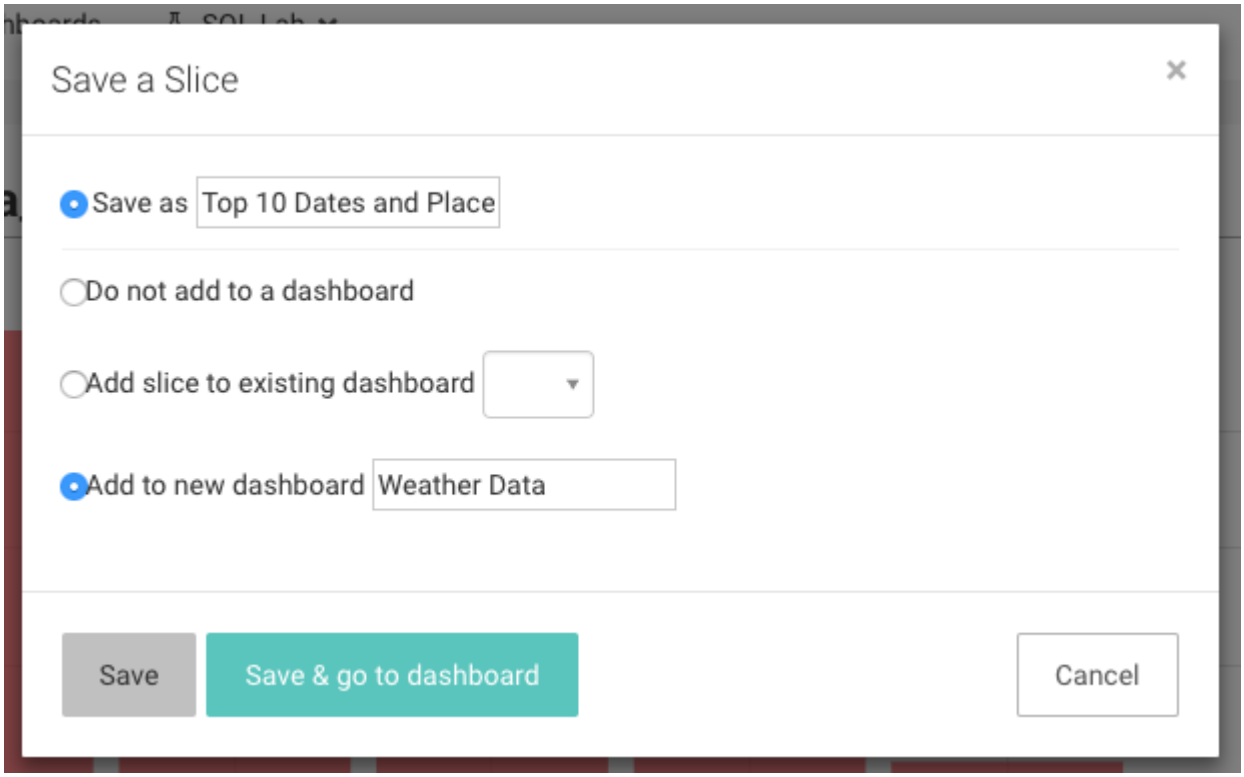
3.5 Creating a slice and dashboard

This view might be interesting to researchers, so let's save it. In Superset, a saved query is called a **Slice**.

To create a slice, click the **Save as** button near the top-left of the explore:

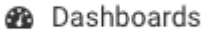


A popup should appear, asking you to name the slice, and optionally add it to a dashboard. Since we haven't yet created any dashboards, we can create one and immediately add our slice to it. Let's do it:



Click Save, which will direct you back to your original query. We see that our slice and dashboard were successfully created:

Let's check out our new dashboard. We click on the **Dashboards** menu.



and find the dashboard we just created.

3.6 Benefits and disadvantages of using Superset

Benefits:

- Supports multiple data sources
- Provides various visualizations (with interactive filters)
- You can create visualizations even without knowledge of SQL
- Visualizations are really beautiful
- Fast reaction to GitHub issues from the team

Disadvantages:

- The tool is rapidly developing so be ready to find bugs and report them to GitHub .
- Problems with customization .

3.7 Security

Security in Superset is handled by Flask App Builder (FAB). FAB is a “Simple and rapid application development framework, built on top of Flask.”. FAB provides authentication, user management, permissions and roles.

3.7.1 Role

Superset ships with a set of roles that are handled by Superset itself. Even though it’s possible for `Admin` users to do so, it is not recommended that you alter these roles in any way by removing or adding permissions to them as these roles will be re-synchronized to their original values as you run your next `superset init` command.

- **Admin**
  Admins have all possible rights, including granting or revoking rights from other users and altering other people's slices and dashboards.
- **Alpha**
  Alpha have access to all data sources, but they cannot grant or revoke access from other users. They are also limited to altering the objects that they own. Alpha users can add and alter data sources.
- **Gamma**
  Gamma have limited access. They can only consume data coming from data sources they have been given access to through another complementary role. They only have access to view the slices and dashboards made from data sources that they have access to. Currently Gamma users are not able to alter or add data sources. We assume that they are mostly content consumers, though they can create slices and dashboards. Also note that when Gamma users look at the dashboards and slices list view, they will only see the objects that they have access to.
- **sql_lab**
  The `sql_lab` role grants access to SQL Lab. Note that while `Admin` users have access to all databases by default, both `Alpha` and `Gamma` users need to be given access on a per database basis.
- **Public**
  It's possible to allow logged out users to access some Superset features.By setting `PUBLIC_ROLE_LIKE_GAMMA = True` in your `superset_config.py`, you grant public role the same set of permissions as for the GAMMA role. This is useful if one wants to enable anonymous users to view dashboards. Explicit grant on specific datasets is still required, meaning that you need to edit the `Public` role and add the Public data sources to the role manually.

3.7.2 Permissions

Roles are composed of a set of permissions, and Superset has many categories of permissions. Here are the different categories of permissions:

- **Model & action**: models are entities like `Dashboard`, `Slice`, or `User`. Each model has a fixed set of permissions, like `can edit`, `can show`, `can delete`, `can list`, and so on. By adding `can delete on Dashboard` to a

role, and granting that role to a user, this user will be able to delete dashboards.

- **Views**: views are individual web pages, like the `explore` view or the `SQL Lab` view. When granted to a user, he/she will see that view in the its menu items, and be able to load that page.
- **Data source**: For each data source, a permission is created. If the user does not have the `all_datasource_access` permission granted, the user will only be able to see Slices or explore the data sources that are granted to them.
- **Database**: Granting access to a database allows for the user to access all data sources within that database, and will enable the user to query that database in SQL Lab, provided that the SQL Lab specific permission have been granted to the user.

4.0 DASHBOARDS

4.1 Introduction

They often provide at-a-glance views of KPIs (key performance indicators) relevant to a particular objective or business process. In the other, "dashboard" has another name for "progress report" or "report."

The "dashboard" is often displayed on a web page which is linked to a database that allows the report to be constantly updated. For example, a manufacturing dashboard may show numbers related to productivity such as number of parts manufactured, or number of failed quality inspections per hour. Similarly, a human resources dashboard may show numbers related to staff recruitment, retention and composition, for example number of open positions, or average days or cost per recruitment

The term dashboard originates from the automobile dashboard where drivers monitor the major functions at a glance via the instrument cluster.

4.2 Research Software Dashboard

The Research Software Dashboard has been produced by RITS's Research Software Development Group to help meet the objective of promoting research software as a first class research output alongside publications and data.

The service provides an inventory where people can discover software written at UCL and developers can share information about their work with colleagues and the wider world

4.3 Features of the service include:

- Integration with IRIS
- so that your software will appear as a research output in your public profile
- Your software can be linked to publications and grants to help show how coding contributes to research
- Search by programming language, libraries or research field to find collaborators or locally-written software you can use
- You decide whether to share information about your software with the world, or just UCL colleagues
- Provide links to online code repositories where you can share your source code and collaborate on new features and fixes
- Get an overview of the state of your software projects including bug reports and feature requests

4.4 Benefits

Digital dashboards allow managers to monitor the contribution of the various departments in their organization. To gauge exactly how well an organization is performing overall, digital dashboards allow you to capture and report specific data points from each department within the organization, thus providing a "snapshot" of performance.

Benefits of using digital dashboards include

- Visual presentation of performance measures
- Ability to identify and correct negative trends
- Measure efficiencies/inefficiencies
- Ability to generate detailed reports showing new trends
- Ability to make more informed decisions based on collected business intelligence
- Align strategies and organizational goals
- Saves time compared to running multiple reports

- Gain total visibility of all systems instantly
- Quick identification of data outliers and correlation

4.5 Classifications

Dashboards can be broken down according to role and are either strategic, analytical, operational, or informational. Strategic dashboards support managers at any level in an organization, and provide the quick overview that decision makers need to monitor the health and opportunities of the business. Das

hboards of this type focus on high level measures of performance, and forecasts. Strategic dashboards benefit from static snapshots of data (daily, weekly, monthly, and quarterly) that are not constantly changing from one moment to the next. Dashboards for analytical purposes often include more context, comparisons, and history, along with subtler performance evaluators. Analytical dashboards typically support interactions with the data, such as drilling down into the underlying details. Dashboards for monitoring operations are often designed differently from those that support strategic decision making or data analysis and often require monitoring of activities and events that are constantly changing and might require attention and response at a moment's notice.

4.6 Types of Dashboards

Digital dashboards may be laid out to track the flows inherent in the business processes that they monitor. Graphically, users may see the high-level processes and then drill down into low level data. This level of detail is often buried deep within the corporate enterprise and otherwise unavailable to the senior executives.

Three main types of digital dashboard dominate the market today: software applications, web-browser based applications, and desktop applications also known as desktop widgets. The last are driven by a widget engine.

Specialized dashboards may track all corporate functions. Examples include human resources, recruiting, sales, operations, security, information technology, project management, customer relationship management and

many more departmental dashboards. For a smaller organization like a startup a compact start up scorecard dashboard tracks important activities across lot of domains ranging from social media to sales.

Digital dashboard projects involve business units as the driver and the information technology department as the enabler. The success of digital dashboard projects often depends on the metrics that were chosen for monitoring. Key performance indicators, balanced scorecards, and sales performance figures are some of the content appropriate on business dashboards

4.7 Dashboards and Scoreboards

- Balanced Scoreboards and Dashboards have been linked together as if they were interchangeable. However, although both visually display critical information, the difference is in the format: Scoreboards can open the quality of an operation while dashboards provide calculated direction. A balanced scoreboard has what they called a "prescriptive" format. It should always contain these components.

- Perspectives – group
- Objectives – verb-noun phrases pulled from a strategy plan
- Measures – also called Metric or Key Performance Indicators (KPIs)
- Spotlight Indicators – red, yellow, or green symbols that provide an at-a-glance view of a measure’s performance.

Each of these sections ensures that a Balanced Scorecard is essentially connected to the businesses critical strategic needs.

The design of a dashboard is more loosely defined. Dashboards are usually a series of graphics, charts, gauges and other visual indicators that can be monitored and interpreted. Even when there is a strategic link, on a dashboard, it may not be noticed as such since objectives are not normally present on dashboards. However, dashboards can be customized to link their graphs and charts to strategic objectives.

4.8 Design

Digital dashboard technology is available "out-of-the-box" from many software providers. Some companies however continue to do in-house development and maintenance of dashboard applications. For example, GEAviation has developed a proprietary software/portal called "Digital Cockpit" to monitor the trends in aircraft spare parts business.

Good dashboard design practices take into account and address the following:

- the medium it is designed for (desktop, laptop, mobile, tablet)
- use of visuals over tabular presentation of data
  - bar charts: to visualize one or more series of data
  - line charts: to track changes in many dependent data sets over a period of time
  - spark lines: to show the trend in a single data set
- use of legends anytime more than one colour or shape is present on a graph
- spatial arrangement: place your most important view on the top left (if the language is written left to right) then arrange the following views in a Z pattern with the most important information following the top-to-bottom, left-to-right pattern
- colour palettes to be colour blind friendly

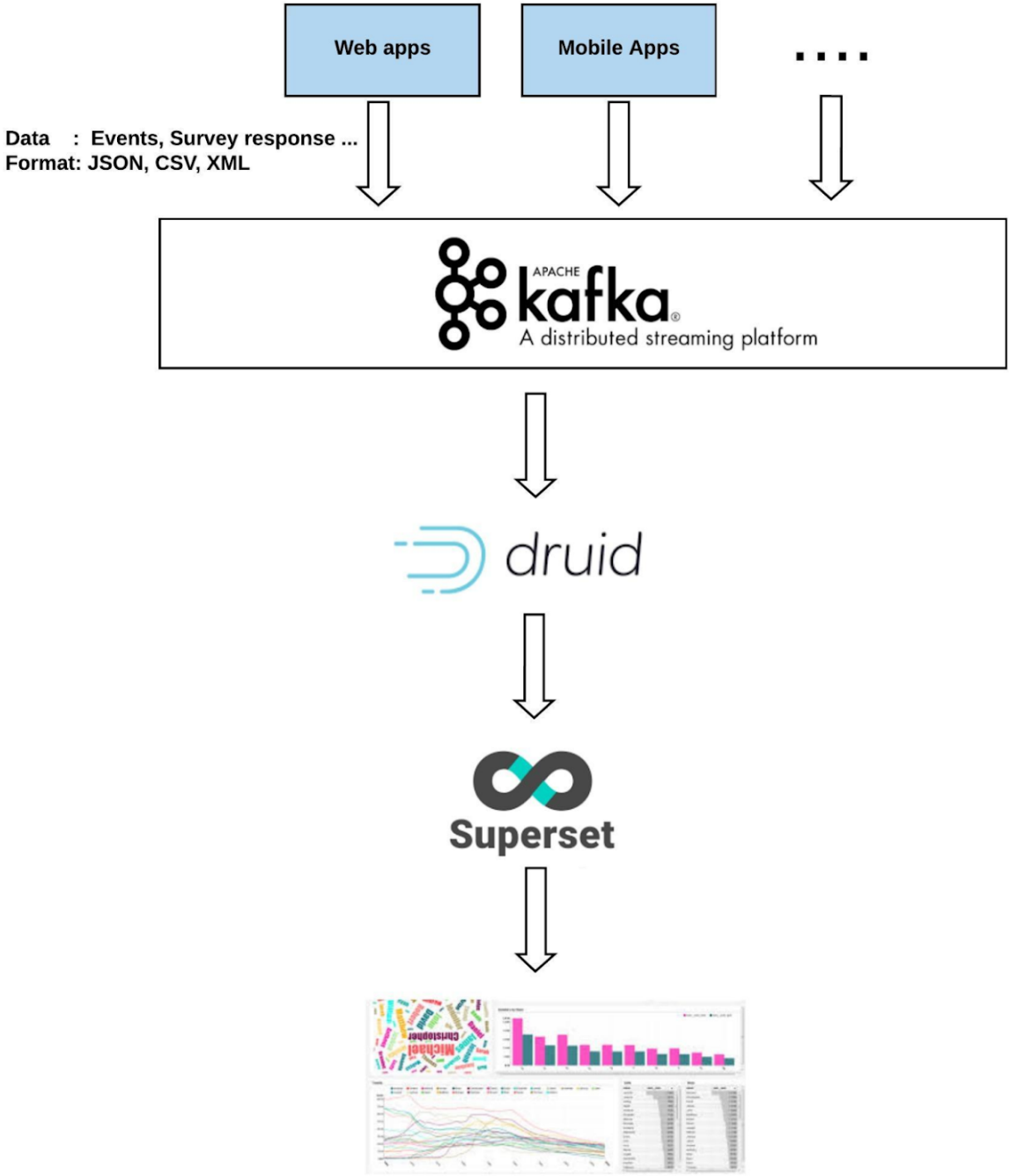
A good information design will clearly communicate key information to users and makes supporting information easily accessible.

4.9 Accessing Quality of Dashboard

There are a few key elements to a good dashboard

1. Simple, communicates easily.
2. Minimum distractions...it could cause confusion.
3. Supports organized business with meaning and useful data.
4. Applies human visual perception to visual presentation of information.
5. It can be accessed easily by its intended audience.

5 Architecture Diagram: Superset supports different types of data tools here we use druid database.



6.0 Overview of CAHPS Surveys

CAHPS surveys ask patients to report on their experiences with care. Several surveys ask about experiences with ambulatory care providers such as health plans, physicians' offices, and mental health plans, while others ask about experiences with care delivered in facilities such as hospitals, dialysis centres, and nursing homes.

7.0 Purpose of the Guide

This Technical Assistance Guide for Analysing Data from the HCBS CAHPS Survey (Data Analysis Guide) is a key component of materials created to help survey sponsors implement the analysis of data and report results from the HCBS CAHPS Survey.

Read the guide to learn:

• What survey-derived information you can report. The surveys include three composite measures about care for adults, five composite measures about care for children, and a rating of the doctor on a scale of 0 to 10.

• How to label and describe the measures. This guide presents the option of a short or long label for each of the CAHPS measures. It also suggests text you can use to describe each measure for those who want more information. All of this text has been tested with consumers to ensure readability and ease of understanding.

• What scores you can report for these measures. The two most suitable methods for calculating scores based on the survey results are "top box" scoring and average scoring, both of which have strengths and weaknesses.

• How to display the scores for consumers. Users of the CAHPS survey can choose among three recommended strategies for displaying results in a way that facilitates understanding and use by consumers.

8.0 Goals of CAPHS Surveys

- Develop Standardized patient questionnaries that can be used to compare results across sponsors and over time.
- Generate tools and resources that Sponsors can use to produce understandable and usable comparative information for both consumers and health care providers.

9.0 Consumer Assessment of Healthcare Providers and Systems

Consumer Assessment of Healthcare Providers and Systems (CAHPS) is a series of patient surveys rating health care experiences in the United States The surveys, conducted annually since 1995, are available in the public domain and focus on healthcare quality aspects that patients find important and are well equipped to assess. Results are used by Medicare in determining Diagnostic Related Group payment for each hospital.

CAHPS surveys are funded and overseen by the Agency for Healthcare Research and Quality (AHRQ), a government organization. The AHRQ does not administer the surveys. Surveys must be administered by a qualified vendor. Vendors become certified by Centers for Medicaid and Medicare Services (CMS).

There are currently several types of CAHPS surveys available. Some of these types include: Hospital (HCAHPS), Clinician Group (CGCAHPS), Hospice (HHCAHPS) and Ambulatory Surgery (ASCAHPS). These surveys may be administered via phone, email or mail distribution, depending on the certification of the vendor administering the survey.

These surveys measure patient experience with various measures that 'should' happen with each medical encounter, such as understandable (https://www.ahrq.gov/cahps/news-and-events/podcasts/measure-patient-experience-podcast.html) communication with doctors, nurses and pain management during a patient's hospital or clinic visit or end of life care.

The three most widely used CAHPS surveys are—

- The CAHPS Health Plan Survey, which asks enrolls in commercial plans, Medicaid, Children's Health Insurance Programs (CHIP), and Medicare about their experiences with health plan services and ambulatory care.
- The CAHPS Clinician & Group Survey (CG-CAHPS), which asks patients to report on their experiences with primary or specialty care received from providers and their staff in ambulatory settings.
- The CAHPS Hospital Survey (HCAHPS), which asks patients about the care delivered during an inpatient stay at a hospital facility.

10.0 CAHPS Surveys

Many of the CMS patient experience surveys are in the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) family of surveys. Others are developed following CAHPS principles and used by CMS but are not CAHPS surveys. All surveys officially designated as CAHPS surveys have been approved by the CAHPS Consortium, which is overseen by the Agency for Healthcare Research and Quality (AHRQ). CAHPS surveys follow scientific principles in survey design and development. The surveys are designed to reliably assess the experiences of a large sample of patients. They use standardized questions and data collection protocols to ensure that information can be compared across healthcare settings. CAHPS surveys are developed with broad stakeholder input, including a public solicitation of measures and a technical expert panel, and the opportunity for anyone to comment on the survey through multiple public comments period through the Federal Register. Finally, many CAHPS measures are statistically adjusted to correct for differences in the mix of patients across providers and the use of different survey modes.

CAHPS surveys are an integral part of CMS' efforts to improve healthcare in the U.S. Some CAHPS surveys are used in Value-Based Purchasing (Pay for Performance) initiatives. These initiatives represent a change in the way CMS pays for services. Instead of only paying for the number of services provided, CMS also pays for providing high quality services. The quality of services is measured clinically, administratively, and through the use of patient experience of care surveys.

10.1 Hospital CAHPS (HCAHPS)

HCAHPS is the first national, standardized, publicly-reported survey of patients' perspectives of hospital care. The HCAHPS Survey (pronounced "H-caps") is a 32-item instrument and data collection methodology for measuring patients' perceptions of their hospital experience. HCAHPS allows valid comparisons to be made across hospitals -- locally, regionally and nationally. The survey was nationally implemented in 2006 and public reporting of hospital scores began in 2008. Since 2012, HCAHPS scores have played a role in hospital payment through the Hospital Value-Based Purchasing program.

10.2 Home Health CAHPS (HHCAHPS)

HHCAHPS (Home Health Care Consumer Assessment of Healthcare Providers and Systems) Survey is the first national standardized and publicly reported survey of home health care patients' perspectives of their skilled home

care. The survey was nationally implemented on a voluntary basis in October 2009, and the survey was required for the Medicare annual payment update requirements beginning with a one-month dry run in the period of July-September 2010, and monthly continuous data collection beginning in October 2010. The National Quality Forum endorsed the HHCAHPS in March 2009.

10.3 Home and Community Based Services Survey

The CAHPS Home and Community-Based Services Survey (HCBS CAHPS) is the first cross-disability survey of home and community-based service beneficiary's experience receiving long-term services and supports. It is designed to facilitate comparisons across the hundreds of state Medicaid HCBS programs throughout the country that target different adults with disabilities, e.g., including frail elderly, individuals with physical disabilities, persons with developmental or intellectual disabilities, those with acquired brain injury and persons with severe mental illness. The HCBS CAHPS Survey is available for voluntary use in HCBS programs as part of quality assurance and improvement activities and public reporting.

10.4 Fee-for-Service (FFS) CAHPS

The Medicare Fee-for-Service (FFS) CAHPS survey collects information from Medicare beneficiaries enrolled in the FFS program. The survey targets a sample of approximately 275,000 beneficiaries. The questions on the survey relate to the enrolee's experience of care with Medicare and their FFS provider.

10.5 Medicare Advantage and Prescription Drug Plan CAHPS (MA and PDP CAHPS)

CMS collects information about Medicare beneficiaries' experiences with, and ratings of, Medicare Advantage (MA-only) plans, Medicare Advantage Prescription Drug (MA-PD) plans, and stand-alone Medicare Prescription Drug Plans (PDP) via surveys of beneficiaries who have been enrolled in their plans for six months or longer. Although all three versions have a nearly identical set of core questions, each version also includes additional questions and response categories related to the enrollees' experiences in their own particular plan type. The health plan survey has been conducted annually since 1998, and the drug plan surveys were added in 2007.

10.6 In-Centre Haemodialysis CAHPS (ICH CAHPS)

ICH CAHPS (In-Centre Haemodialysis Consumer Assessment of Healthcare Providers and Systems) Survey is the first national, standardized, and publicly-reported survey of patients' perspectives of dialysis care; it is also the first CAHPS survey focusing solely on a chronic disease. ICH CAHPS measures perceptions of patients with End-Stage Renal Disease (ESRD) receiving life-sustaining in-Center haemodialysis care. In 2007, the National Quality Forum (NQF), an organization established to standardize health care quality measurement and reporting, formally endorsed measures from the ICH CAHPS.

10.7 Hospice Experience of Care Survey

The CAHPS® Hospice Survey gathers information on the experiences of hospice patients and their informal caregivers' perspectives of their loved ones' care with hospice services. Current trends are toward increased use of hospice services in the U.S. More than a million Americans are receiving hospice services annually. The Hospice CAHPS® Survey started national implementation in January 2015.

10.8 CAHPS for ACOs

The CAHPS for ACOs survey was developed to collect information about patient experience and care received from Accountable Care Organizations (ACOs) participating in the Medicare Shared Savings Program (Shared Savings Program) and the Pioneer ACO Model. ACOs are groups of doctors, hospitals, and other health care providers, who come together voluntarily to give coordinated high-quality care to their Medicare patients. The Shared Savings Program rewards ACOs that lower their growth in health care costs while meeting performance standards on quality of care and putting patients first. The CAHPS for ACOs survey was first implemented in 2012.

10.9 Outpatient and Ambulatory Surgery CAHPS (OAS CAHPS)

The Consumer Assessment of Healthcare Providers and Systems Outpatient and Ambulatory Surgery Survey (OAS CAHPS) collects information about patients' experiences of care in hospital outpatient departments (HOPDs) and ambulatory surgery Centres (ASCs). As of December 2012, there were approximately 5,357 Medicare-certified ASCs and about 3,360 HOPDs. In 2006, there were almost 35 million ambulatory surgical visits; approximately 20 million surgeries and procedures were performed in HOPDs, and approximately 15 million were

performed in ASCs. Medicare payments to ASCs have increased by 24% over the years, from $2.9 billion in 2007 to $3.6 billion in 2012. Considering the growing number of ASCs and the increase in Medicare expenditures for outpatient surgical services in both ASCs and HOPDs, the implementation of OAS CAHPS will provide statistically valid data from the patient perspective to inform quality improvement and comparative consumer information about outpatient facilities.

10.10 CAHPS for PQRS

The CAHPS for PQRS survey was developed to collect information about patient experience and care within medical group practices participating in the Physician Quality Reporting System (PQRS). The Physician Quality Reporting System (PQRS) is a quality reporting program that encourages individual eligible professionals (EPs) and group practices to report information on the quality of care they provide to Medicare. The CAHPS for PQRS survey was first implemented in 2013 to large group practices (with 100 or more eligible professionals) that registered for the PQRS Group Practice Reporting Option (GPRO) and used the Web Interface to report their quality measures. Implementation of the survey has expanded in subsequent years to include smaller group practices and group practices using GPRO methods other than the Web Interface.

10.11 CAHPS for MIPS Survey

Overview: The Merit-based Incentive Payment System (MIPS) is one track of the Quality Payment Program, where clinicians earn a performance-based payment adjustment to their Medicare payment. Clinicians participating in MIPS have the flexibility to choose the measures and activities that are most meaningful to their practice to demonstrate performance. The CAHPS for MIPS survey is an optional quality measure that groups participating in MIPS can elect to administer. The CAHPS for MIPS survey is also an optional improvement activity that groups can attest to administering.

11.0 Principles of CAHPS Survey Development

The development of CAHPS surveys is governed by a set of principles meant to ensure that the surveys are scientifically sound and provide information that is specific, understandable, and actionable. These principles influence survey content and design, survey administration, and issues around the reporting of results. In keeping with these principles, CAHPS surveys—

- Ask about aspects of care for which a patient or enrolee is the best or only source of information.
- Ask about the aspects of care that patients say are most important.
- Ask patients to report on the health care they receive.
- Reflect input from a broad spectrum of stakeholders, including patients, clinicians, administrators, accrediting bodies and policymakers.
- Build on existing research and available tools.
- Are standardized to ensure that data collection, analyses and reports are consistent across all users of a given survey.

12.0 Uses of CAPHS Survey

Once AHRQ's CAHPS Consortium releases a CAHPS survey into the public domain, it is available for any organization to use for its own purposes. The results of CAHPS surveys are typically used to monitor and drive improvements in patient experience with care and to better inform consumers about health care providers in their area. Some organizations incorporate the survey results into programs that reward or recognize health care providers for providing high-quality care.

Key users of the CAHPS surveys include the following:

- **The Centres for Medicare & Medicaid Services (CMS).**
  For many of the CAHPS surveys, CMS is the organization that sponsors the national implementation of the survey, sets the policies for survey administration, analyses the data, and publishes the results in private and public reports (including Web sites such as Hospital Compare and Physician Compare). In some cases, such as HCAHPS, CMS also uses the survey scores along with other quality measures to help determine payment incentives that reward high performing health care providers.

- **The National Committee for Quality Assurance (NCQA).**
  NCQA incorporates CAHPS Health Plan Survey results into its accreditation process and its national performance reports. It also uses a version of the Clinician & Group Survey for its Patient-Centred Medical Home (PCMH) Recognition Program
- **Veterans Health Administration (VHA) and Department of Defence (DOD).**
  The VHA and DOD use versions of the Clinician & Group Survey and HCAHPS to assess patient experience in both military and privately contracted clinics and hospitals across the country.
- **Multistake holder organizations.**
  The CG-CAHPS and HCAHPS surveys have become part of the portfolio of performance measures used by local, regional, and State collaborative organizations committed to improving primary, specialty, and inpatient care in their communities.

13.0 CAHPS Clinician & Group Survey

- **Patient Survey Results**

  Patients answered a survey about their experiences at their doctors' offices. (If you display a summary score, add:) The information shown here combines their answers into one score

- **Getting Timely Appointments, Care, and Information**

  The survey asked patients how often they got appointments for care as soon as needed and timely answers to questions when they called the office.

- **Doctor–Patient Communication**

  (12-month version): The survey asked patients how often their doctors explained things clearly, listened carefully, showed respect, provided easy-to-understand instructions, knew their medical history, and spent enough time with the patient.

  (Visit version): The survey asked patients if their doctors explained things clearly, listened carefully, showed respect, provided easy to understand instructions, knew their medical history, showed respect, and spent enough time with the patient.

- **Helpful, Courteous, and Respectful Office Staff**

(12-month version): The survey asked patients how often office staff were helpful and treated them with courtesy and respect.

(Visit version): The survey asked patients if office staff were helpful and treated them with courtesy and respect.

- **Doctor's Attention to Your Child's Growth and Development**

  The survey asked parents if the doctor talked about their child's growth, behaviours, moods and emotions, and ability to learn and get along with others.

- **Doctor's Advice on Keeping Your Child Safe and Healthy**

  The survey asked parents if the doctor talked about keeping their child from getting injured, the food the child eats, physical activity, and household problems.

- **Patients' Rating of Doctor**

  The survey asked patients to rate their doctors on a scale of 0 to 10, with 0 being the worst and 10 being the best.

14.0 Creating Your Analysis Dataset

Before conducting analyses with your data, you need to carry out several tasks to prepare the data:

• Task 1: Code and enter the data.

• Task 2: Conduct an audit.

• Task 3: Identify and exclude ineligible cases.

• Task 4: Clean the data.

• Task 5: Identify and include only complete and partial completed surveys.

• Task 6: Recode variables for analysis.

Task 1: Code and Enter the Data

The first task is to enter the raw data into an electronic data file where each row or case in the data file represents a returned survey and each column represents a question. The responses to each question must use the code numbers, or pre -codes, contained in the questionnaires.

Task 2: Conduct an Audit

If survey data are collected using paper questionnaires (whether mailed or administered by telephone), the second task is to audit a small random sample of the entered data by comparing hard-copy forms with the results of data entry.

This step enables you to catch any systematic errors that show up consistently across all questionnaires.

Task 3: Identify and Exclude Ineligible Cases

Several situations render a case ineligible for analysis. One common situation that should be coded as ineligible is when the respondent reports that he or she has not visited the sampled entity (e.g., a physician or medical group). This would be indicated by a "no" response to Question 1 (e.g., "Our records show that you got care from the provider named below in the last 6 months. Is that right?"). Other situations that should be coded as ineligible for analysis include the following, with exceptions noted:

• Deceased. If the individual has died between the visit(s) and receipt of the questionnaire, the case should be coded as ineligible.

• Language barrier. If the potential respondent has a language barrier, the instrument is not available in the respondent's preferred language, and no one was available to translate the questions for the respondent, the case should be coded as ineligible.

• Proxy. In some settings of Care it may be acceptable for someone to answer the questions on behalf of the target respondent (proxy) and be included in the analysis. In these situations, a case- mix variable indicating whether the data were self-reported, reported with assistance, or reported on behalf of the target respondent may be included when analysing the data. In most situations, however, proxy data should not be included in the analysis or in the numerator when calculating the survey response rate.

• Child Pediatric patients (under 18 years old) are ineligible from the adult version of a survey unless the child is emancipated.

Task 4: Clean the Data

In many cases, the data set you have created will have imperfections even after pre-testing programs and conducting an audit. Please take the following steps to check for and fix these imperfections before reporting any results. The cleaning of data is less imperative with Web surveys and optical scanning because most of these problems cannot happen with these data collection methods. Nevertheless, one should ensure against these problems in every mode.

Check for out-of-range values

Out-of-range responses occur when respondents provide inappropriate responses for a question.

• If the value was due to a data entry error, fix it.

• If the value truly was an out-of-range response, set the out-of-range value to "missing," which drops it from the data analysis for that particular question.

Check for skip pattern problems

• If the respondent answered the screener question such that the subsequent dependent items should be missing, but the items are answered, the subsequent dependent items should be set to missing.

• If the screener question is left blank, and the subsequent dependent items are answered, the responses are retained, and the screener question should remain as missing.

Check for duplicates

Duplicates can occur if the respondent responds via two data collection modes (e.g., Web and telephone, or Web and mailed questionnaire) and one mode arrives at the same time or soon after the other and the case slips through the receipt control system, or if there are errors in data entry (the same questionnaire is entered or scanned twice). If two questionnaires are received for the same respondent, keep the first questionnaire that is received and remove the duplicate unless the questionnaire received later has substantially more completed items than the first questionnaire.

Task 5: Identify Complete and Partial Completed Surveys for Inclusion

Once the data are cleaned, the next step is to code questionnaires as complete, partially complete, or incomplete and then remove the incomplete cases from the data file. Only complete and partially completed questionnaires should be retained for the analysis dataset.

• Complete questionnaire: A questionnaire is considered complete if responses are available for at least half of the key survey items and at least one reportable item.

• Partially completed questionnaire: A questionnaire is considered partially complete if responses are available for at least one reportable item, but less than half of the key items.

• Incomplete questionnaire: A questionnaire is incomplete if the individual did not answer at least one reportable item.

Task 6: Recoding Variables for Analysis

This section outlines the types of variables found in the CAHPS surveys and discusses what variables may need to be recoded for analysis.

Types of Variables in the CAHPS Surveys

The following are the variable types for most of the core items in CAHPS surveys:

Variable type Sample Scale

Min – Max Response Values

Dichotomous 1= Yes

2= No

0 – 1

Global rating 0 = Worst to 10 = Best 0 – 10

3-point response scale 1 = Yes, definitely

2 = Yes, somewhat 3 = No

1 – 3

4-point response scale 1= Never

2 = Sometimes 3 = Usually 4 = Always

1 – 4

Recoding Variables

Although not required, it is best to recode certain variables for ease in analysis, reporting, and interpretation. When recoding variables, always save the original version of that variable. The types of precodes that require recoding are presented below.

Dichotomous Variables. For easier interpretation of the results, the "positive" response should have the highest value. Raw data for this type of variable will need to be recoded as the precodes in CAHPS surveys typically set the values of the responses to 1=Yes and 2=No rather than 0 and 1 where 1=Yes or the highest category. Below is sample SAS code for recoding; the logic can be applied in any software package.

Variables with Three-Point Response Scales. In some cases, reverse coding may be necessary to ensure that the most positive response has the highest value—for example, where "Yes, definitely" is the most positive response. The precodes have the most positive category as the lowest value.

Negatively Worded Items. Reverse coding is also required for some of the supplemental items where the question is negatively worded – that is, responding negatively to the question (e.g., Never) is the positive response.

Identifying the Level of Analysis and Reporting

CAHPS survey data are often collected at three levels: respondents, entities.Therefore, the results can also be examined at each of these levels.

• Respondent level: A respondent is defined as an individual who has completed or partially completed a CAHPS survey. Respondent-level survey results are calculated across all respondents in the database. If the dataset includes multiple entities, respondent-level results would ignore the association of the respondent with a given entity.

• Entity level: Entity level results are calculated across respondents within each given entity. When results are calculated this way, and there is more than one entity, the results can be case-mix adjusted to account for different patient characteristics within each entity that might affect scores.

• Group level: Group-level results are calculated across respondents within a given group. When there is more than one group, results can be case-mix adjusted to account for different patient characteristics within each group that might affect scores.

Calculating Frequencies

Frequencies indicate the number and percentage of respondents answering each response option for the items on the survey. The frequencies can be calculated in two ways: excluding missing values from the calculation of percentages or including them.

Calculating Top Box and Other Proportional Scores

Top box scores are generally calculated as the percentage of respondents who chose the most positive option(s) on a given response scale (e.g., "Yes, definitely" or "Always"). The main advantage of using top box scores in reporting is that it is easier to explain and interpret one score for providers, staff, and consumers. A disadvantage is that top box scores only provide information for one end of the response scale; users have no idea of the distribution of responses across the remaining options.

Calculating proportional scores for an individual survey item: A proportional score is calculated by aggregating results across respondents for the reporting unit or entity.

Calculating top box and other proportional scores for a composite measure: The scores for a composite measure are equal to the average or mean of the proportion of responses (excluding missing data) in each response category across the items in the composite. The following steps show how proportions for composite measures are calculated:

• Step 1 – Calculate the proportion of cases in each proportional score category for each question in a composite.

• Step 2 –Calculate the average proportion responding to each category across the questions in the composite.

Calculating Mean or Average Scores

The mean or average score for an item is calculated by taking the mean across all response categories. For example, in a sample of 10 respondents, if 4 out of 10 respondents answered “Always” or “4” on the response scale, 2 answered “Usually” or a “3”, 2 answered “Sometimes” or “2”, and “2” answered “Never” or “1,” the mean would be 2.8 out of 4 (4+4+4+4+3+3+2+2+1+1 /10). The top box score for this measure would be 40%.

15.0 Choosing a Display Strategy

This section discusses the optional use of a summary score and recommends a few ways to display the results of the CAHPS Clinician & Group Survey:

• Table with word icons

• Table with graphic symbols

• Bar charts

In offering these alternative approaches, the authors recognize that not everyone is ready or willing to highlight differences in physician- or practice-level performance. That said, when selecting a display strategy, we encourage you to focus on the message you want consumers in your audience to take away from the information and how you want them to use the information. If your display does not support those messages or uses, you may need to consider the alternatives.

16.0 CONCLUSIONS:

- This document introduces a great visualization tool, Superset, and how to use it to build dashboard out of a SQLite database. However, I only scratched the surface of Superset; there are so many more you can do with it.

- This study provides support for reliability and validity of the CAHPS Item Set for Addressing Health Literacy. These items can serve to assess whether healthcare providers have communicated effectively with their patients and as a tool for quality improvement.

17.0 REFERENCES:

https://en.wikipedia.org/wiki/Consumer_Assessment_of_Healthcare_Providers_and_Systems

https://www.ahrq.gov/sites/default/files/wysiwyg/cahps/surveys-guidance/cg/about/cg_3-0_overview.pdf

https://www.ahrq.gov/cahps/bibliography/index.html

http://www.dhcs.ca.gov/dataandstats/reports/Documents/MMCD_Qual_Rpts/CAHPS_Reports/CA2012-13_CAHPS_Summary_Report_F3.pdf

https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/CAHPS/OAS-CAHPS.html

https://en.wikipedia.org/wiki/Druid

https://en.wikipedia.org/wiki/Apache_Kafka

https://superset.incubator.apache.org/

https://medium.com/@InDataLabs/superset-benefits-and-limitations-of-the-open-source-data-visualization-tool-by-airbnb-8dc8ac81efa9

https://en.wikipedia.org/wiki/Time_series_database

https://en.wikipedia.org/wiki/Online_analytical_processing

https://sherpasoftware.com/blog/structured-and-unstructured-data-what-is-it/https://en.wikipedia.org/wiki/Apache_Hadoop

https://en.wikipedia.org/wiki/NoSQL

https://spring.io/understanding/NoSQL

https://trailhead.salesforce.com/en/modules/reports_dashboards/units/reports_dashboards_visualizing_data