

G20 Energy End-Use Data and Energy Efficiency Metrics initiative

Rethinking data collection amid C19 crisis

29 Oct 2020



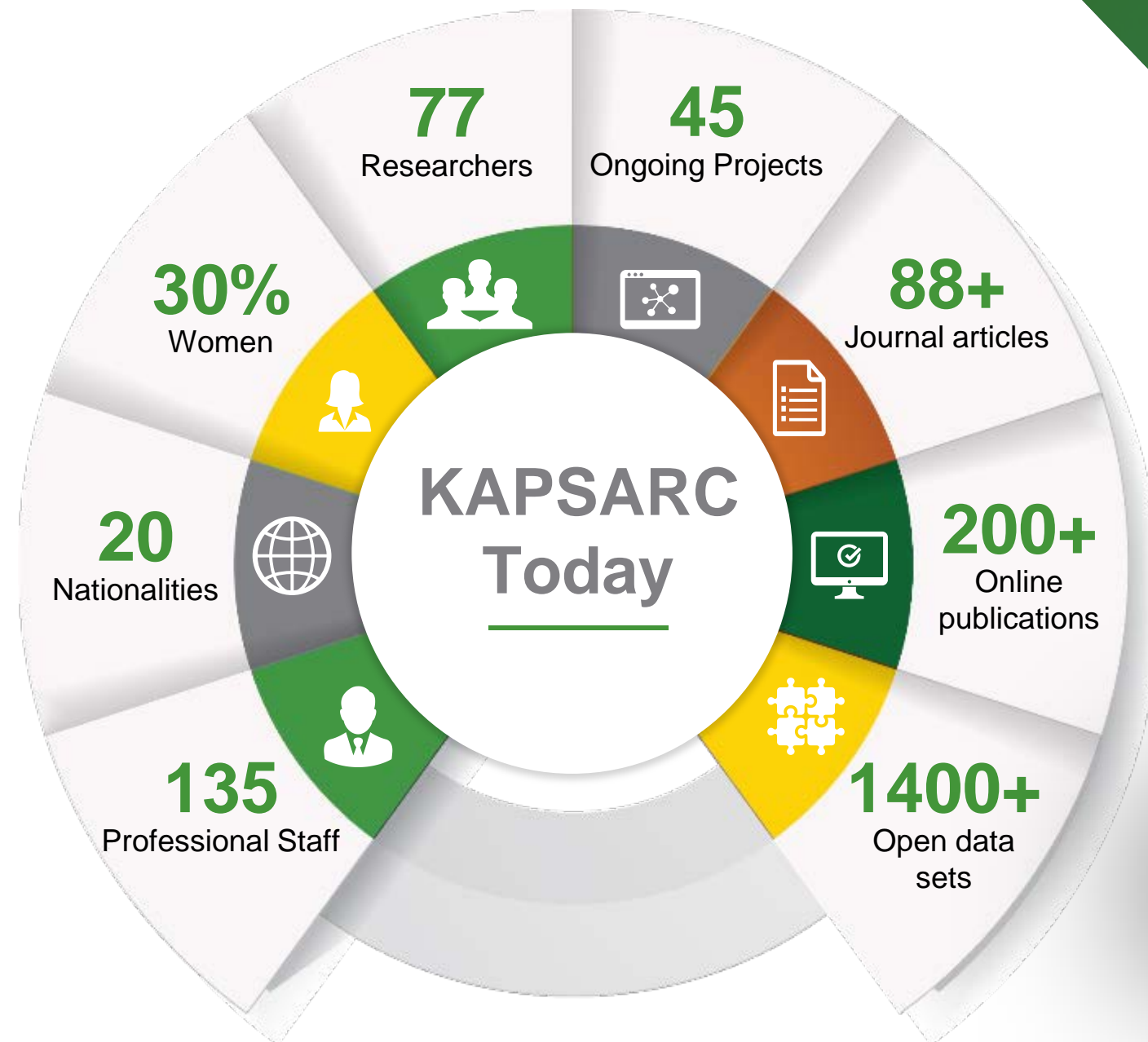
Agenda

- Data challenge: Regional data availability for researchers and energy modelers
- Our approach to solve: OPEN Data-portal and Data-hub that are API (Application Programming Interface) ready
- Data Quality & Energy Models Initiative
- Examples of new datasets and dashboards: Eg. Covid 19, Google mobility data, Energy economics indicators

Amar Amarnath, Head of Energy Information Management - KAPSARC

Climate and Environment
Policy and Decision Science
Energy and Macroeconomics
Energy Transitions and Electric Power
Markets and Industrial Development
Transport and Urban Infrastructure
Energy Information Management

Co-leading T20 Saudi Arabia
Think20 (T20)



Overview

The King Abdullah Petroleum Studies and Research Center (KAPSARC) is a non-profit global institution dedicated to independent research into energy economics, policy, technology and the environment.

Mission

KAPSARC's mandate is to advance the understanding of energy challenges and opportunities facing the world and Saudi Arabia, through objective research that informs quality decision-making.

Data Challenges & Consequences

- ❑ Historical data availability
- ❑ Granular sectoral data needs
 - Energy consumption and prices by sector and customer type
 - Sectoral investments, Sectoral employment, Wages, Govt spending,
- ❑ Short-Spanned data for macroeconomic indicators
- ❑ Unavailability of high frequency data
- ❑ Data/reporting formats are not machine readable
- ❑ Incorrect representation of economic relationships
- ❑ Difficult to understand revisions / changes upstream
- ❑ Unable to represent sectoral level and granular relationships
- ❑ Poor representation of economic linkages
- ❑ Unable to perform policy analysis and projections
- ❑ Delay in model updates, upgrades

Attribute Suppression

The removal of an entire part of data in a dataset.

Record Suppression

The removal of an entire record in a dataset.

Character Masking

The change of the characters of a data value – it is typically partial.

Pseudonymization

The replacement of the identifying data with made up values.

Generalization

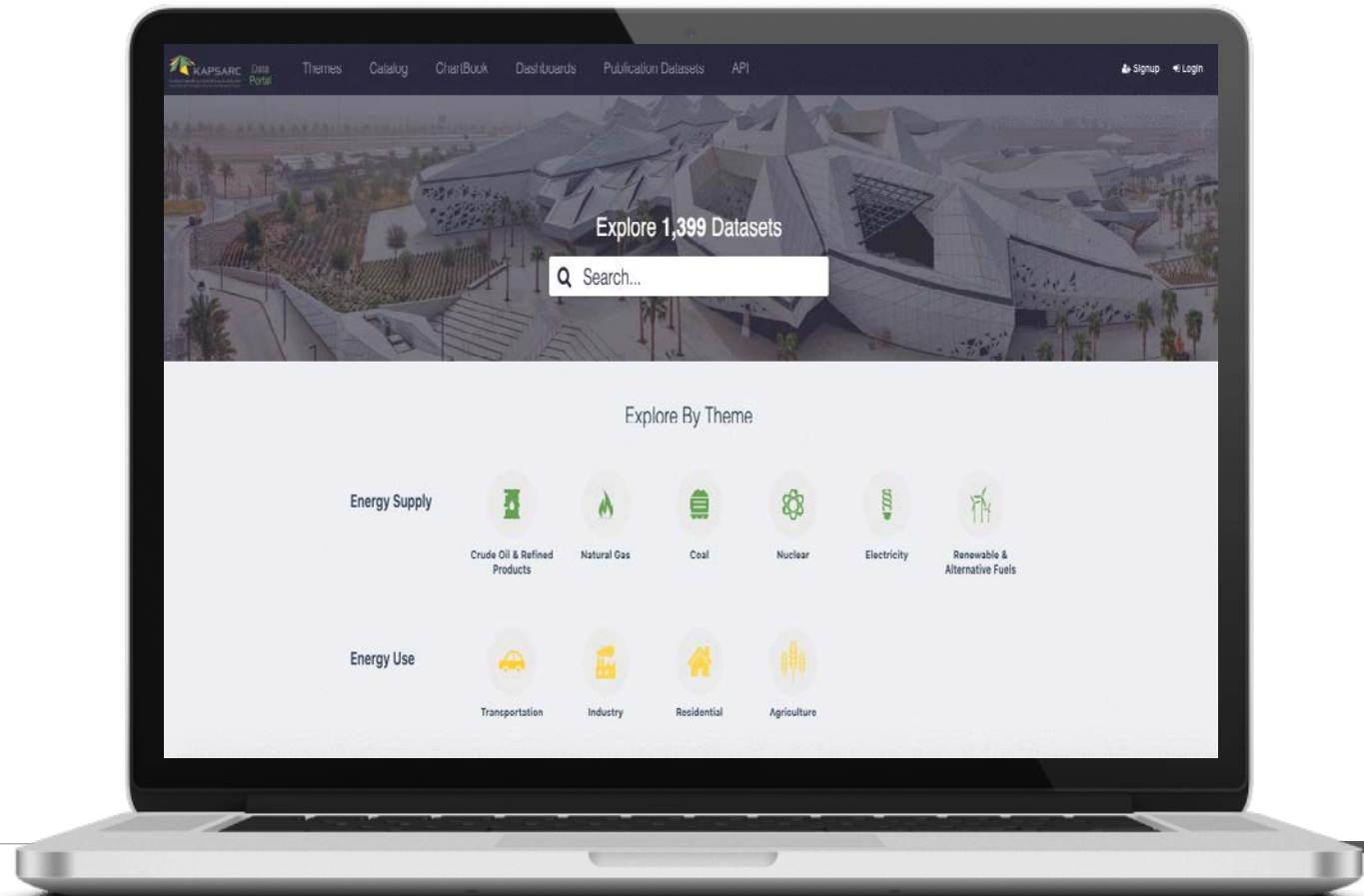
A deliberate reduction in the precision of data.

Data Aggregation

Converting a dataset from a list of records to summarized values.

#1 KAPSARC Open Data -portal

- KAPSARC launched open data portal to public in 2016, grown to around 1400 public datasets.
- The portal is designed to enable users to better understand energy, economy and environment.
- Critical energy economic data is available in an easy to use machine readable format.



Datasets classified into 16 themes

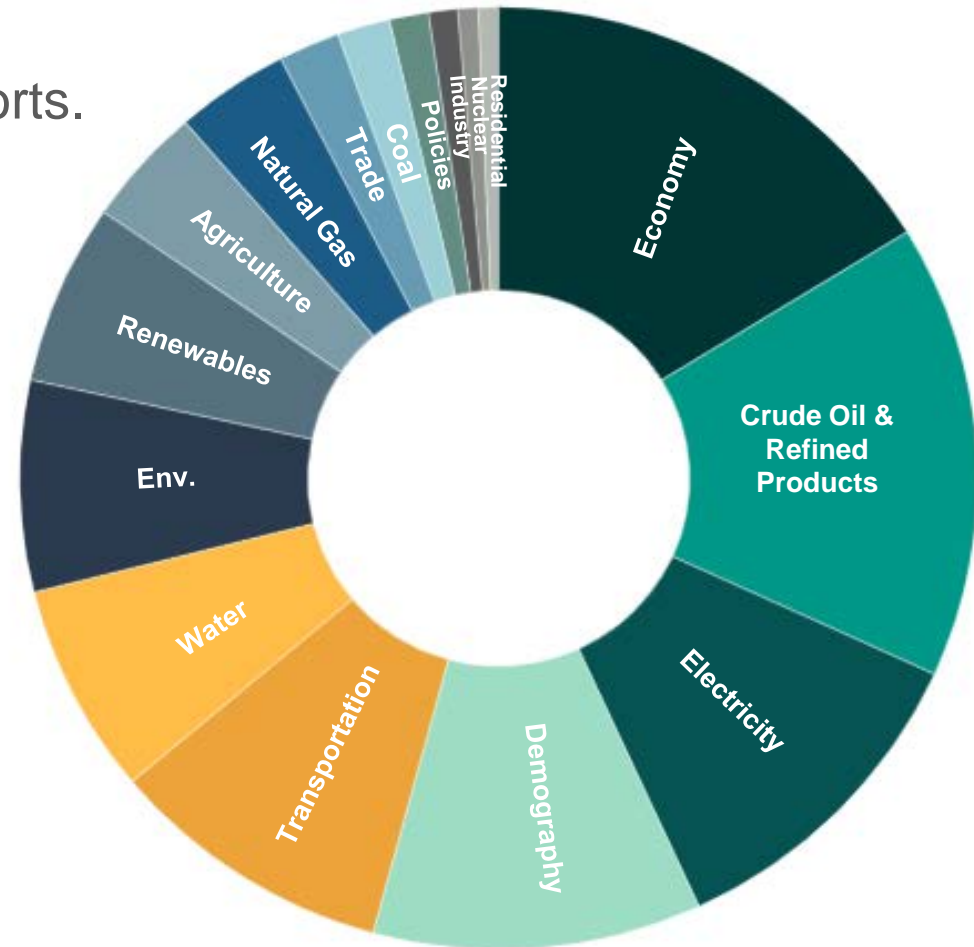
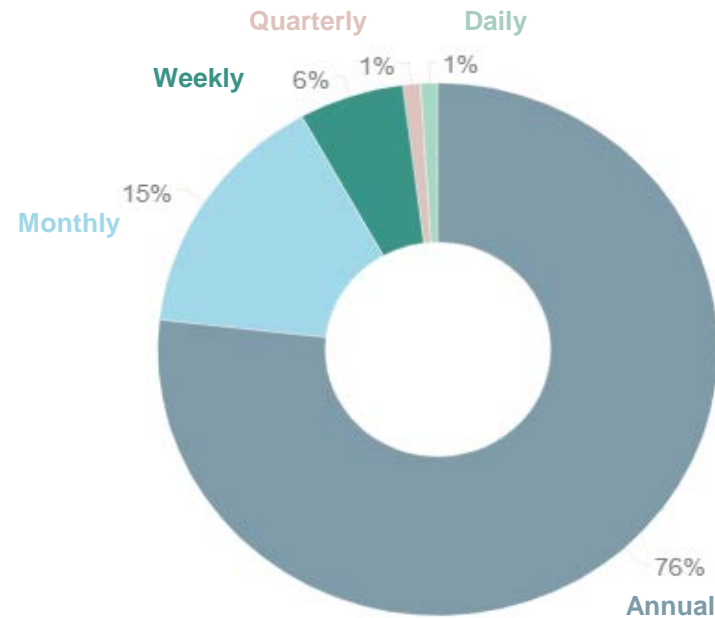
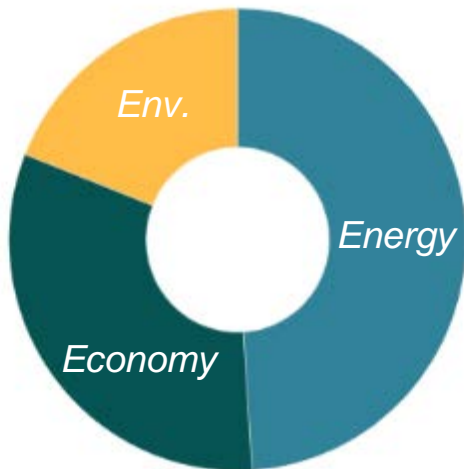
Themes

16

Datasets

1400

- Clean and machine-readable energy economy datasets
- User can search, filter, create charts, maps.
- Data also can be extracted through API or customized exports.

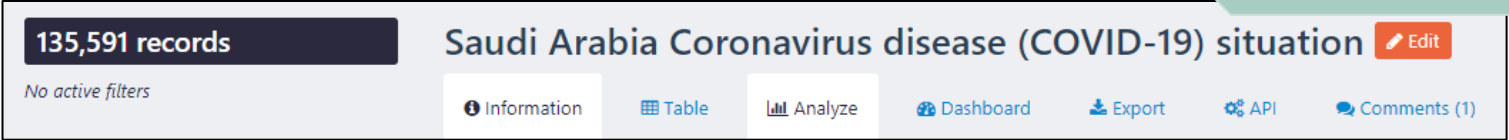


Countries

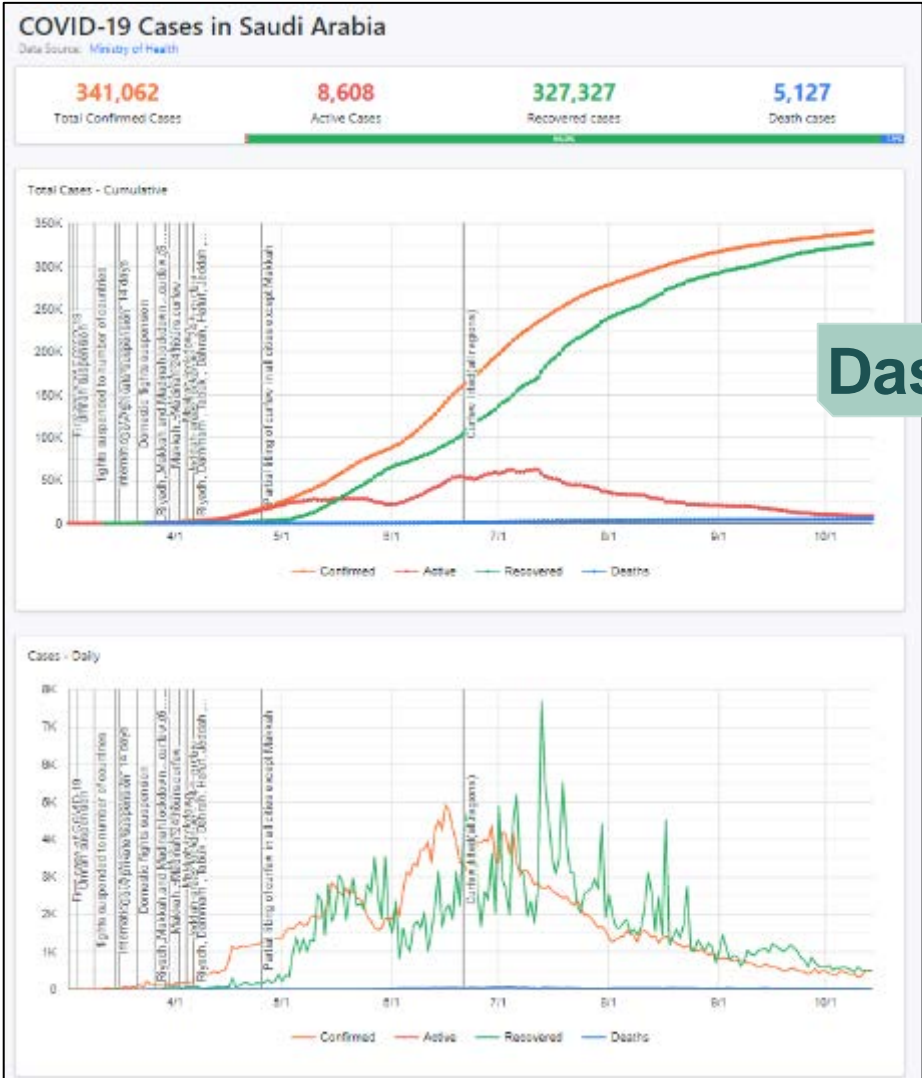
30+

COVID19 dataset and dashboard

Dataset



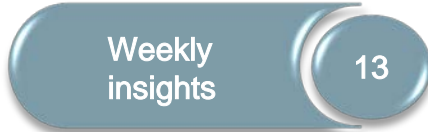
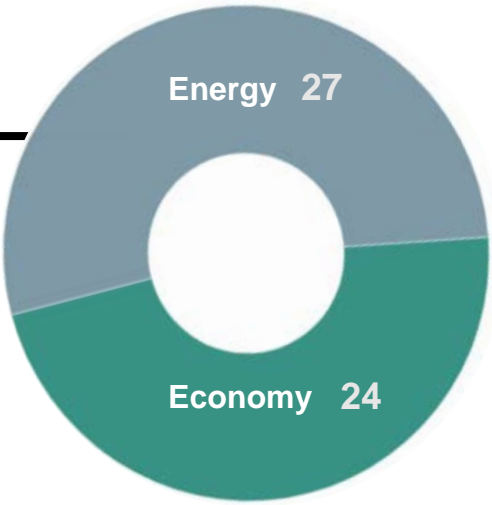
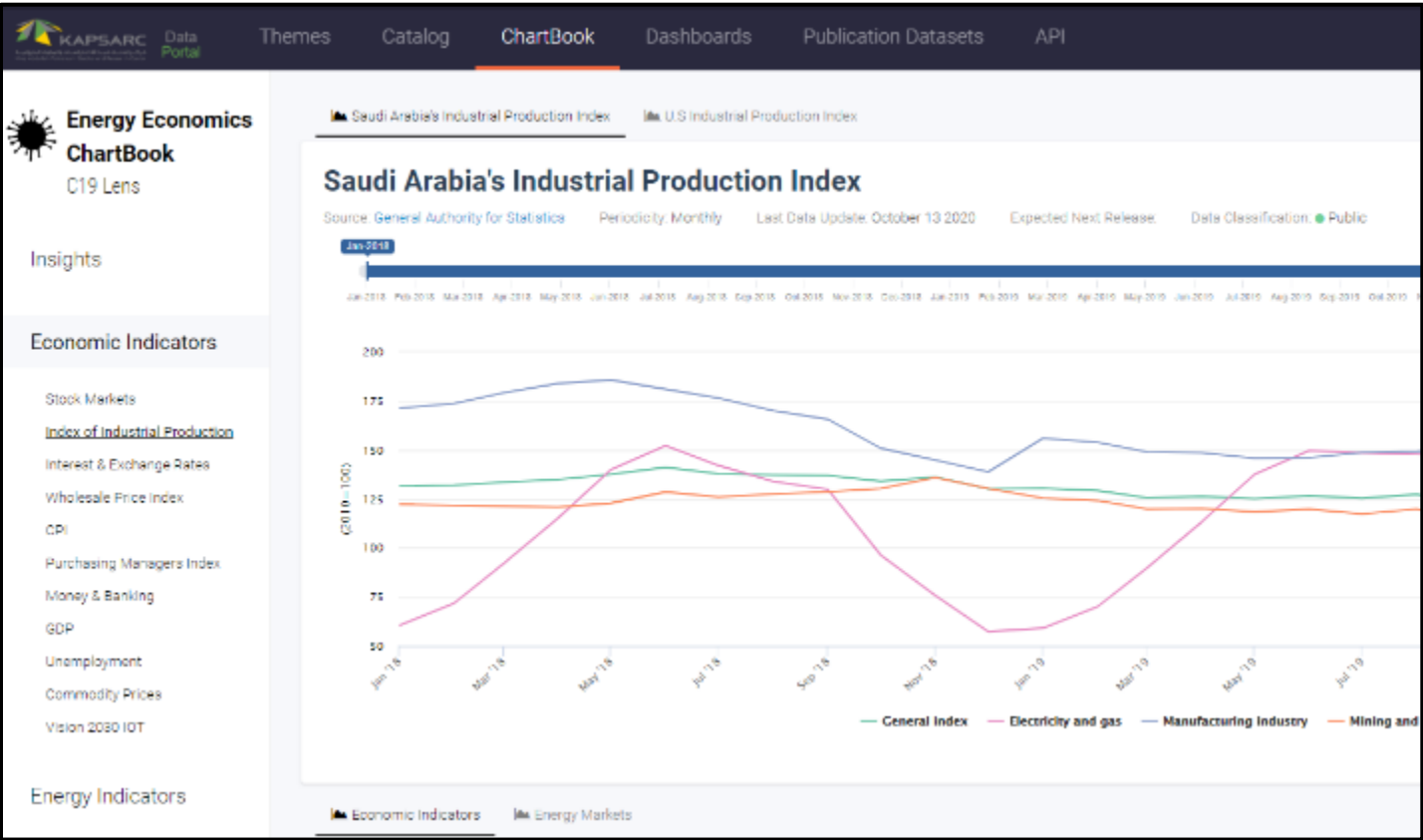
Of downloads



Dashboard

News








Economic indicators

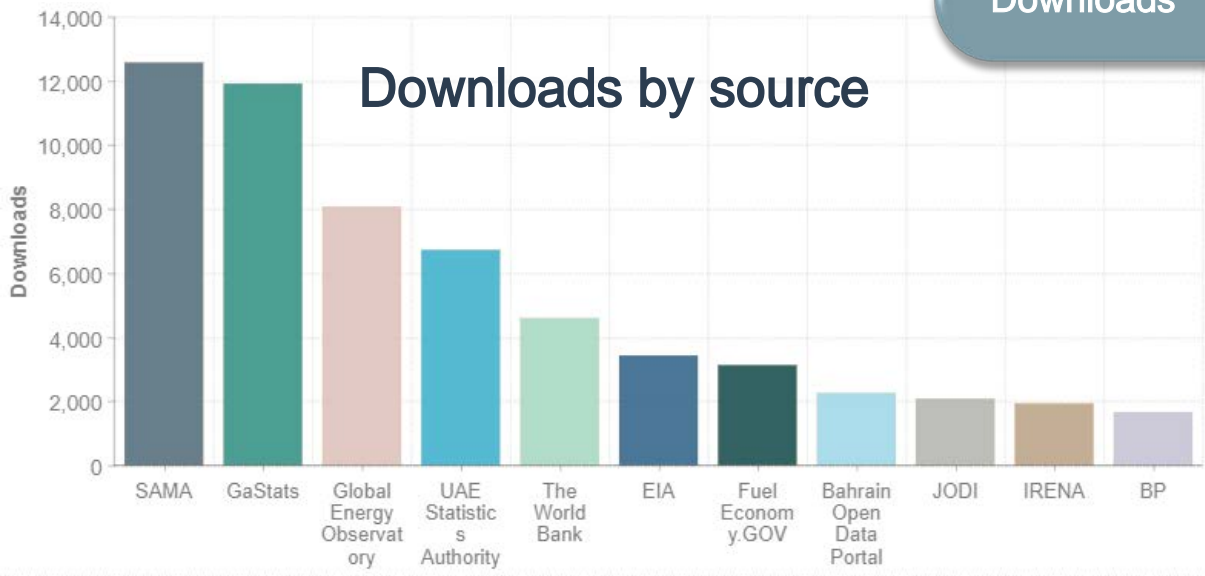
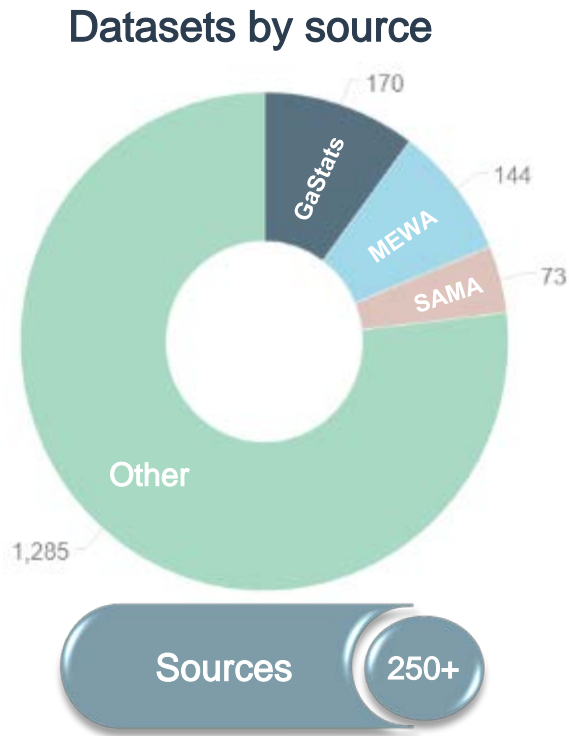
- Industrial production index
- Interest rate
- Price indices
- GDP
- Unemployment



Energy indicators

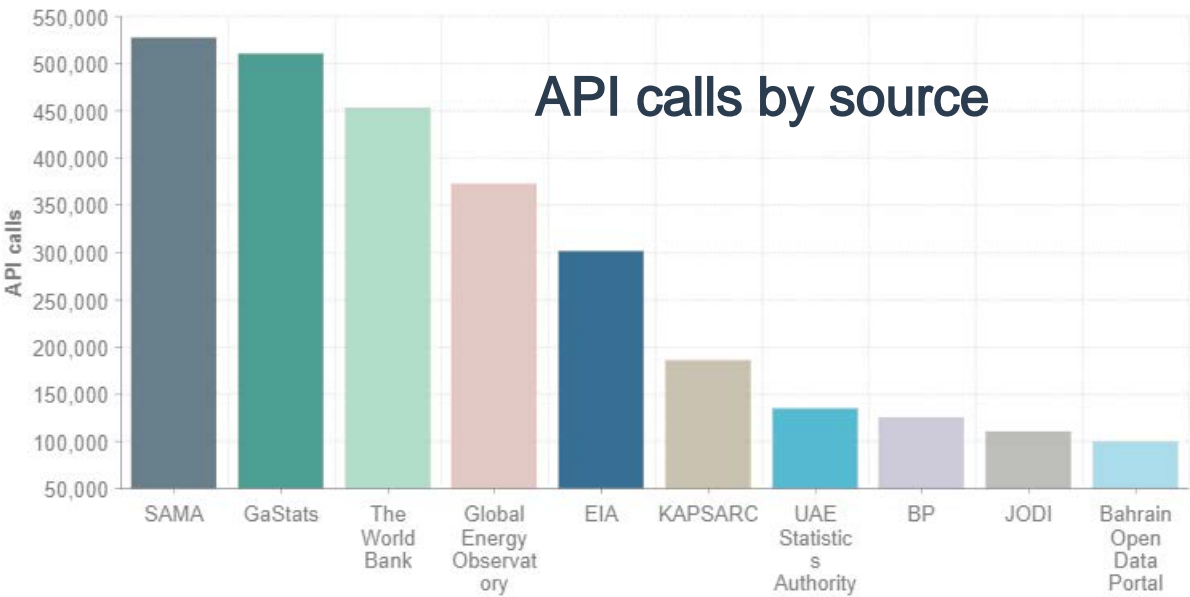
- Prices
- Production
- Consumption
- Trade
- Electricity

Data sources represented in data portal



Downloads 148K

API calls 3.7M



Dataset Identifier	saudi-arabia-coronavirus-disease-covid-19-situation
Downloads	62,221
Last checked date	August 31, 2020
Predecessor	Ministry of Health https://covid19.moh.gov.sa/
Country	Saudi Arabia
ISO region	APAC
Data classification	Public
Related datasets	https://datasource.kapsarc.org/explore/dataset/saudi-arabia-covid-19-situation
Unit of measure	Persons
Publisher Periodicity	Daily
Themes	Demography, KAPSARC Insights
Keywords	COVID-19
License	Public Domain
Language	English
Modified	October 18, 2020 3:52 PM
Publisher	King Abdullah Petroleum Studies and Research Center
Reference	https://datasource.kapsarc.org/pages/eechartbook/
Attributions	Saudi Arabia Coronavirus disease (COVID-19) situation

Tools for informed policymaking

We offer free access to KAPSARC's data sets and research tools, to assist policymakers and to advance the understanding of energy economics and environment policy worldwide.

Data and Tools



GO TO

Publications

Workshops

Data / Tools

SEARCH

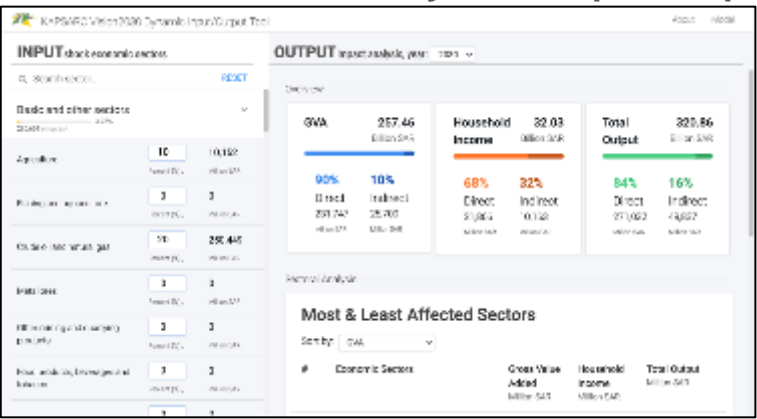
Try "Oil research"

Search

Energy Policy Scenario Model Tools

20 open web applications/tools are published on our website showcasing the policy scenario models serving as a platform to better understand energy

KAPSARC Vision 2030 Dynamic Input-Output Tool



2020

- #17 Global energy relations vis.
- #18 KOSeMOSYS energy optimization
- #19 V2030 dynamic I/O tool v.2
- #20 GVAR market shock simulator

2019

- #13 Fuel substitution calculator
- #14 Datahub for global models
- #15 V2030 dynamic I/O tool v.1
- #16 Energy policy simulator

2018

- #10 Web iKTAB for Behavioral Analysis
- #10 KGEMM
- #11 Data Insights app
- #12 Marine transport analysis

2017

- #5 Saudi Arabia building energy efficiency tool.
- #6 India renewable energy policy atlas
- #7 Data profiler
- #9 KAPSARC Energy Model

2016

- #3 Vehicle Fleet Model
- #4 KAPSARC Data Portal

2015

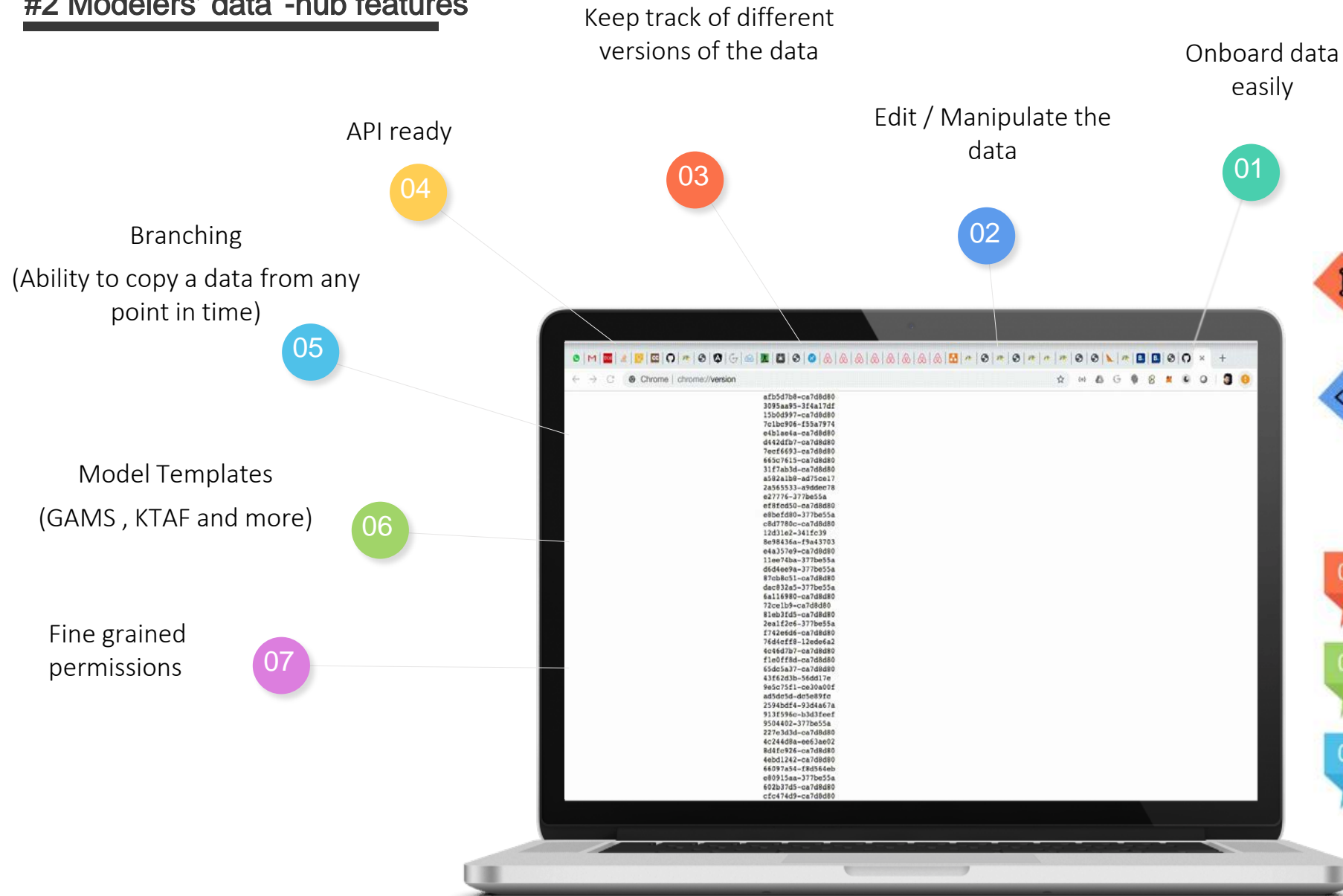
- #1 China Energy Policy Atlas
- #2 Solar PV ToolKit

<http://apps.kapsarc.org/kbeat/>



Saudi Arabia Building Energy Efficiency Tool

#2 Modelers' data -hub features



KOSeMOSYS Open Source Energy Model is an open source model for energy planning. It has been developed to provide energy system planners with a tool to evaluate the impact of energy efficiency retrofits on the energy system and to find the optimal level of investments in energy efficiency retrofits. Energy efficiency retrofits are measures that can be implemented in buildings, industry and transport to reduce energy consumption and greenhouse gas emissions.



5 C's of data quality

Currency

- Deliver new and updated content in a timely manner

44%

Automated datasets

Correctness

- How accurate is the data?
- Does it contradict other trusted resources?
- Check discrepancies
- Report to modelers and publishers.

32%

From PDF files

Completeness

- Are all the data fields populated as per your format/ requirement
- Metadata completeness

94%

Of metadata

Coverage

- Getting everything researcher needs to derive insights from data
- Coverage analysis
- Benchmark analysis

	Saudi Arabia	Kuwait	UAE	Qatar	Bahrain	Oman	China	India
Population								
Population , Population by age / gender								
Numebr of Briths , Crude Birth rate (per 1,000 people)								
Number of Deaths , Crude Death rate (per 1,000 people)								
Fertility rate, total (births per woman)								
Life expectancy at birth								
Mortality rate								
Rural and Urban population								
Population Density (persons/sq.km)								
Migration								
Population Projections								
Labor								
Age dependency ratio (% of working-age population)								
Labor force, Labor force by age / gender								
Unemployment , Unemployment (% of total labor force)								
Wages								
Employees by economic activity or Occupation								
Number of weeks of maternity leave								
Average Working Hours								
Education								
Literacy , Literacy by age/ gender , Literacy rate								
School enrollment								
Higher education statistics								

Consistency

- Standardization of identifiers and cross references

	Saudi Arabia	Kuwait	UAE	Qatar	Bahrain	Oman	China	India
Population								
Population , Population by age / gender								
Numebr of Briths , Crude Birth rate (per 1,000 people)								
Number of Deaths , Crude Death rate (per 1,000 people)								
Fertility rate, total (births per woman)								
Life expectancy at birth								
Mortality rate								
Rural and Urban population								
Population Density (persons/sq.km)								
Migration								
Population Projections								
Labor								
Age dependency ratio (% of working-age population)								
Labor force, Labor force by age / gender								
Unemployment , Unemployment (% of total labor force)								
Wages								
Employees by economic activity or Occupation								
Number of weeks of maternity leave								
Average Working Hours								
Education								
Literacy , Literacy by age/ gender , Literacy rate								
School enrollment								
Higher education statistics								

Rapid pace of change in “data supply chain”

Hyper-speed

Digital services to grow 100x faster than today

Hyper-scale

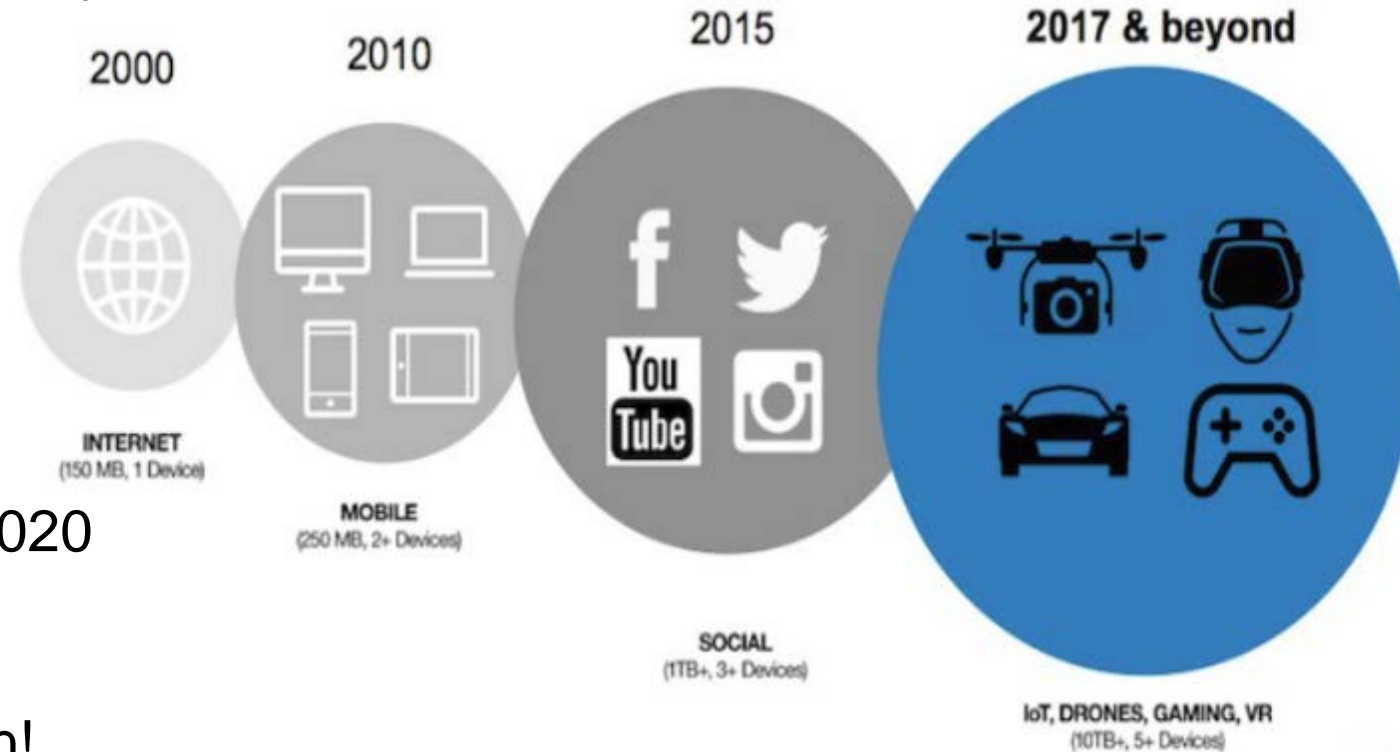
Digital apps, next 4 years ~ last 40 years

Hyper-connected

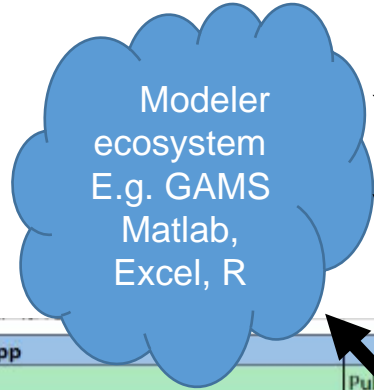
30 billions of edge devices connected in 2020

~Two megabytes every second per person!

Source:IDC



Modelers' data-hub vs open data-portal



<http://apps.kapsarc.org/datahub>

<https://datasource.kapsarc.org/>

Modeler1 input

- dataset1
- dataset2
- ...



Store modelers
input closed
data repository

Opendata input

- dataset1
- dataset2
- ...



Master open data
repository

Model Output

- Modeler1
 - Modeler2
 - ...
1. Data versions
 2. Model versions
 2. Recipes
 3. Charts
 4. API ready



Compare &
cross
walk model
outputs

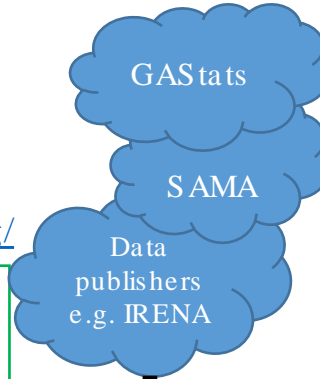
1400 datasets
80M records

OpenKAPSARC
Datasets

- Oil
- Gas
- ...



1. Meta data
2. Search
3. Chart
4. Maps
5. Export
6. API ready



Model / App	Status
KAPSARC Energy Model - KSA	Published
KAPSARC Energy Model - GCC	Published
KAPSARC Global energy macroeconomic model	Published
KAPSARC Toolkit for behavioral analysis (KTAB)	Published
Saudi Arabia Energy Policy Simulator	Published
KAPSARC Vehicle fleet model	Published
KAPSARC Building Energy Assessment Tool	Published
KAPSARC India renewable energy policy atlas	Published
KAPSARC Solar photovoltaic toolkit	Published
KAPSARC Crude Oil Storage Cost	Published
Data Profiler	Published
KAPSARC Datahub	Published
Vision 2030 Dynamic Input-Output Tool	Published
KAPSARC Transport Analysis Framework (KTAF)	Published
KAPSARC OSEMOSSYS model - KSA	Under Dev / EIM
KAPSARC GVar model - KSA	Under Dev / EIM
Carbon Trading Blockchain Registry Toll	Under Dev / EIM
KAPSARC world energy dependency	Under Dev / EIM
KAPSARC Upstream model of investment decision options	Under Dev/Modeler
KAPSARC Oil Market Outlook	Under Dev/Modeler
KAPSARC GCC Gas Model	Under Dev/Modeler
KSA Power Systems model	Under Dev/Modeler
KSA Hourly Price Forecast Model	Under Dev/Modeler
KAPSARC Residual Supplier Toolkit	Under Dev/Modeler

Key take away

- Request machine readable granular open data
- Share data transformation recipes
- Partnership with data publishers to enhance data quality and granularity
- Request user comments on the data portal to assess data quality
- Suggest datasets or data sources through the data portal

Let's partner on OPEN

Data

Models

Tools

Policy Pathway Insights

The screenshot shows the bottom portion of a web application. At the top right, there is a dark navigation bar with links for 'About', 'Suggest', and 'العربية'. Below this is a main navigation bar with 'Catalog' (highlighted with an orange underline), 'ChartBook', 'Dashboards', 'Publication Datasets', and 'API'. On the right side of this bar are 'Signup' and 'Login' links. The main content area has a title 'Saudi Arabia Coronavirus disease (COVID-19) situation'. Below the title is a row of interactive options: 'Information', 'Table', 'Analyze', 'Dashboard', 'Export', 'API', and 'Comments (1)'. The 'Comments (1)' option is circled in red. In the top right corner of the main content area, there is a dark box with the text 'Let's partner on OPEN' and a list of categories: 'Data', 'Models', 'Tools', and 'Policy Pathway Insights'. At the bottom right, there are social media icons for Twitter, Facebook, LinkedIn, and Email.

About Suggest العربية

Catalog ChartBook Dashboards Publication Datasets API

Signup Login

Saudi Arabia Coronavirus disease (COVID-19) situation

Information Table Analyze Dashboard Export API Comments (1)

Let's partner on OPEN

Data

Models

Tools

Policy Pathway Insights

Backup, technical slides

Research objects needs to be interoperable – 21 Rs



scientific methods - reproducible, repeatable, replicable, reusable



access – referenceable, retrievable, reviewable



understanding – replayable, reinterpretable, reprocessible



new use – recomposable, reconstructable, repurposable



social – reliable, respectful, reputable, revealable



curation – recoverable, restorable, reparable, refreshable

A publication is not the scholarship itself,

The actual scholarship is the complete data, code and instructions.

- Stanford's Jon Claerbout

De Roure, D. 2014. The future of scholarly communications. Insights: the UKSG journal, 27, (3), 233-238. DOI 10.1629/2048-7754.171

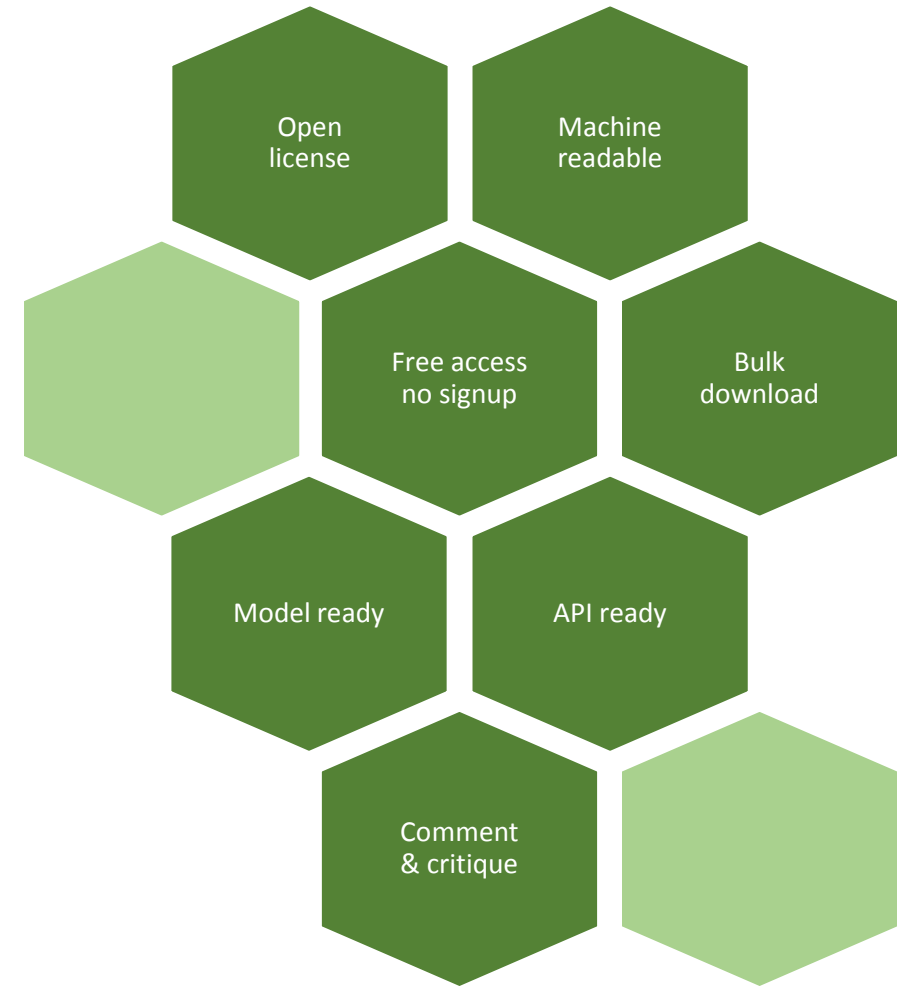
Is data transparent?

Shared data plays a crucial role in progress, and is reused in unexpected ways bringing greater good

- ☐ Is open licensed?
- ☐ Is machine readable?
- ☐ Is free of charge? – no login?
- ☐ Is bulk download ready?
- ☐ Is up-to-date?
- ☐ Is data recipe open?
- ☐ Is meta data attached?
- ☐ Is social? – can share, like, comment, critique?

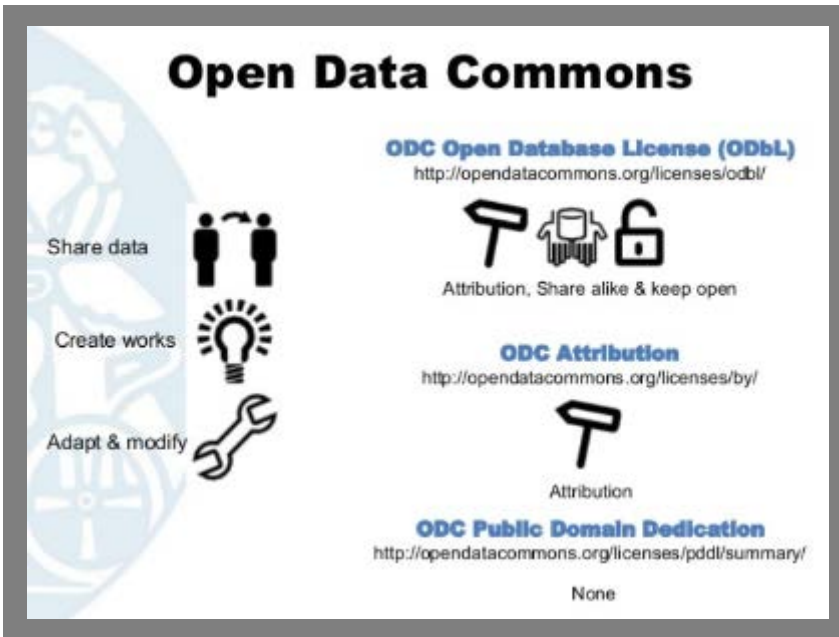
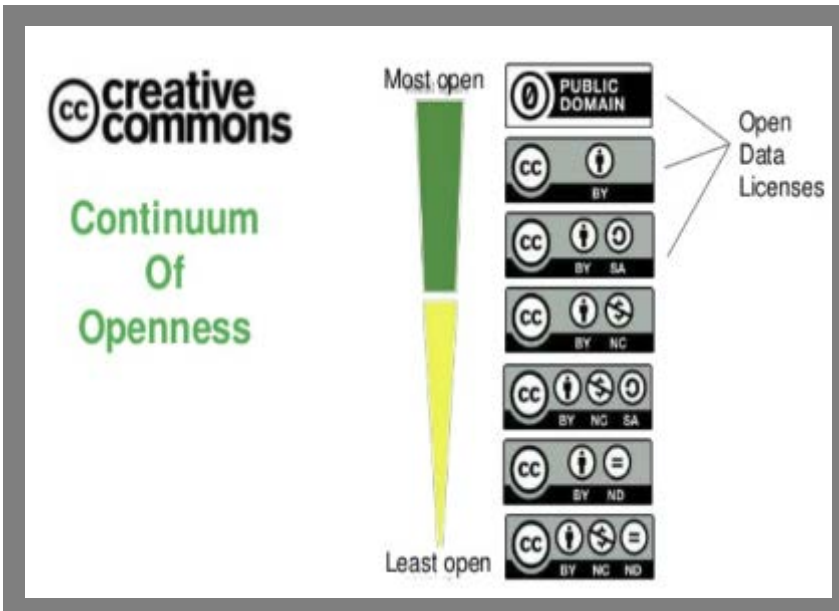
Common Task Framework in 1960s
data share enabled today's
artificial intelligence NLP development

Word Error Rate %	Model	Year Reference
49	GMM	1995
19	GMM	2013
17	DNN	2011
14	DNN	2014
13	DNN	2013
13	Deep RNN	2014
10	Conv DNN	2014




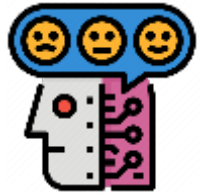


Transforming via open access to research inputs-outputs

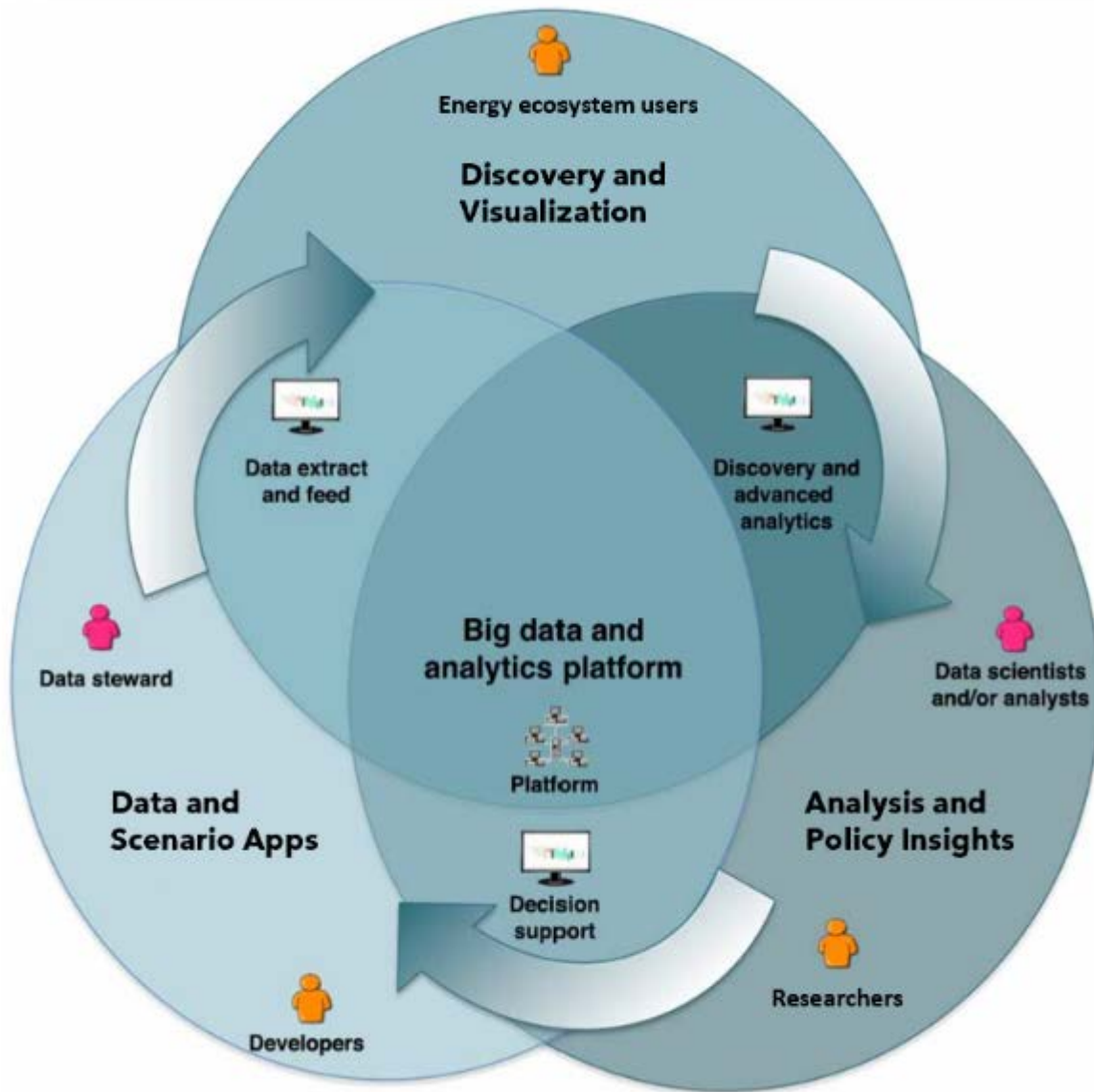
- ❑ Data resolution and access
 - ❑ Sector, regional and monthly granularity
 - ❑ Future will pivot on API/IOT data streams
- ❑ Enable discussion threads on data, add easy to share data web links and web widgets, make data social
- ❑ Specify “data use” license; share, create, adapt, attribute
 - ❑ Creative commons has 7 configurations
 - ❑ Public domain, Open data license (attribute, share alike, no commercial, no derivative)
 - ❑ Open data commons has 5 configurations
 - ❑ ODBL - attribute, share alike, keep open, create work, adapt
 - ❑ ODC - attribute
 - ❑ ODC Public domain dedication - none



Research data use-cases that need advanced BigData platforms

	Research Use-case	Platform capability
	Policy relevant event data processing of multi-language news sources and social media data	Context analytics: Natural language processing, text analytics, translation, speech to text e.g. Amazon comprehend, dataiku
	Minute resolution power system data for forecasting demand, operational planning in the electricity sector	Realtime data-stream analytics, IOT data into metrics and alerting platforms e.g. Prometheus , Thanos , Apache NIFI , EMR and Dataiku for predictions
	GPS and mobility data to analyze travel behavior and demand	Parsing google mobility data, data pipelines and big data and spatial data analytics e.g. Nightlight data with ESRI , EMR , Dataiku
	Country and subnational level sentiment analysis on climate and environment-related topics	Context analytics: Text analytics, translation, sentiment analytics e.g. twitter data into Dataiku , Amazon Comprehend

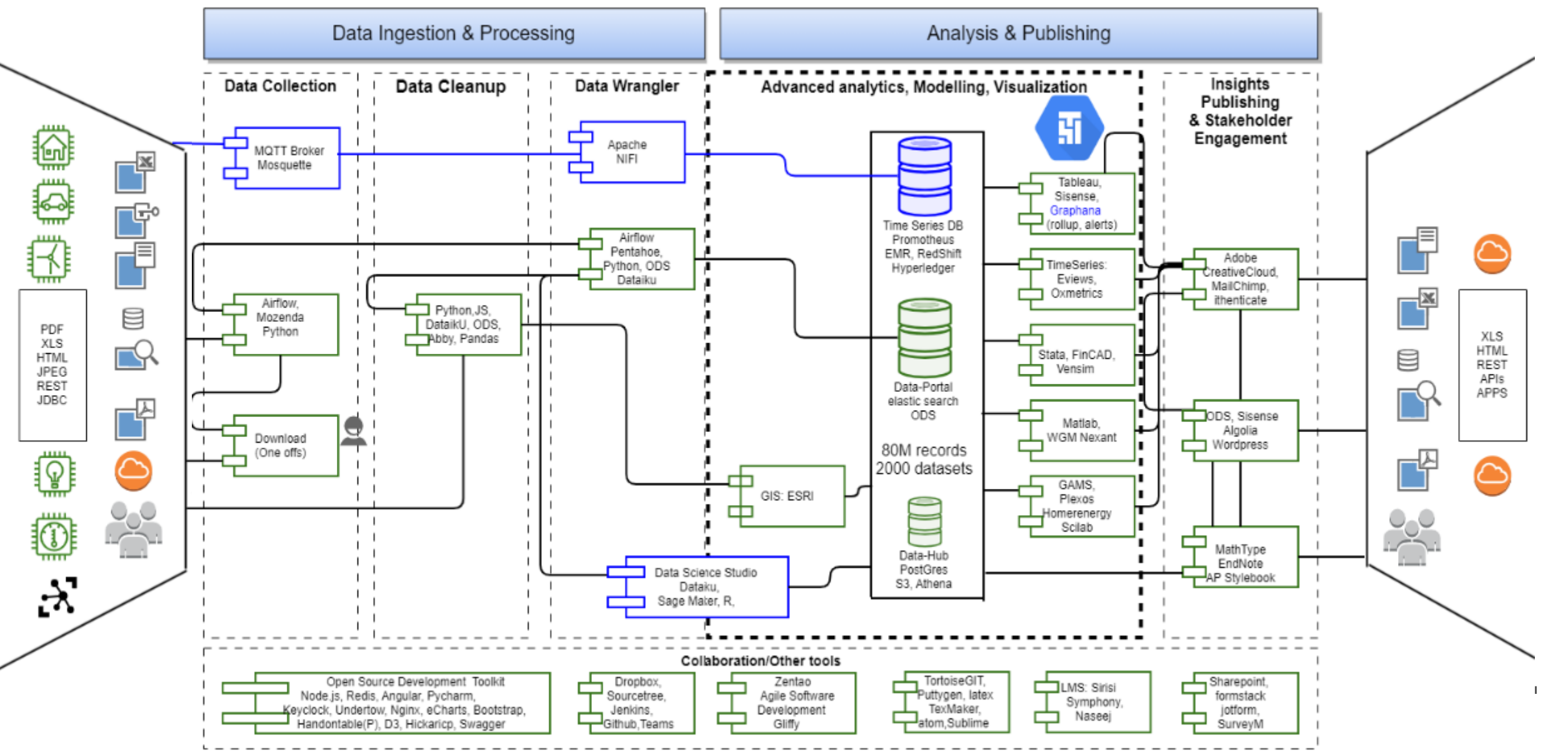
Big data capability: “point of departure” and “point of arrival”



Capability	Point of Departure	Point of Arrival (Gaps)
Data Repo	Elastic Search, S3 Datahub, RDS	+ Hyperledger
Data Engineering	Airflow, Python, DataScienceStudio Dataiku	+ Amazon Elastic Map Reduce, Athena NiFi, Prometheus
Energy and Macro Economics modeling and simulation framework for policy insights	GAMS Eviews Matlab Plexos	+ Sage Maker
Data Science Studio Business Intelligence	DSS Dataiku Sisense	

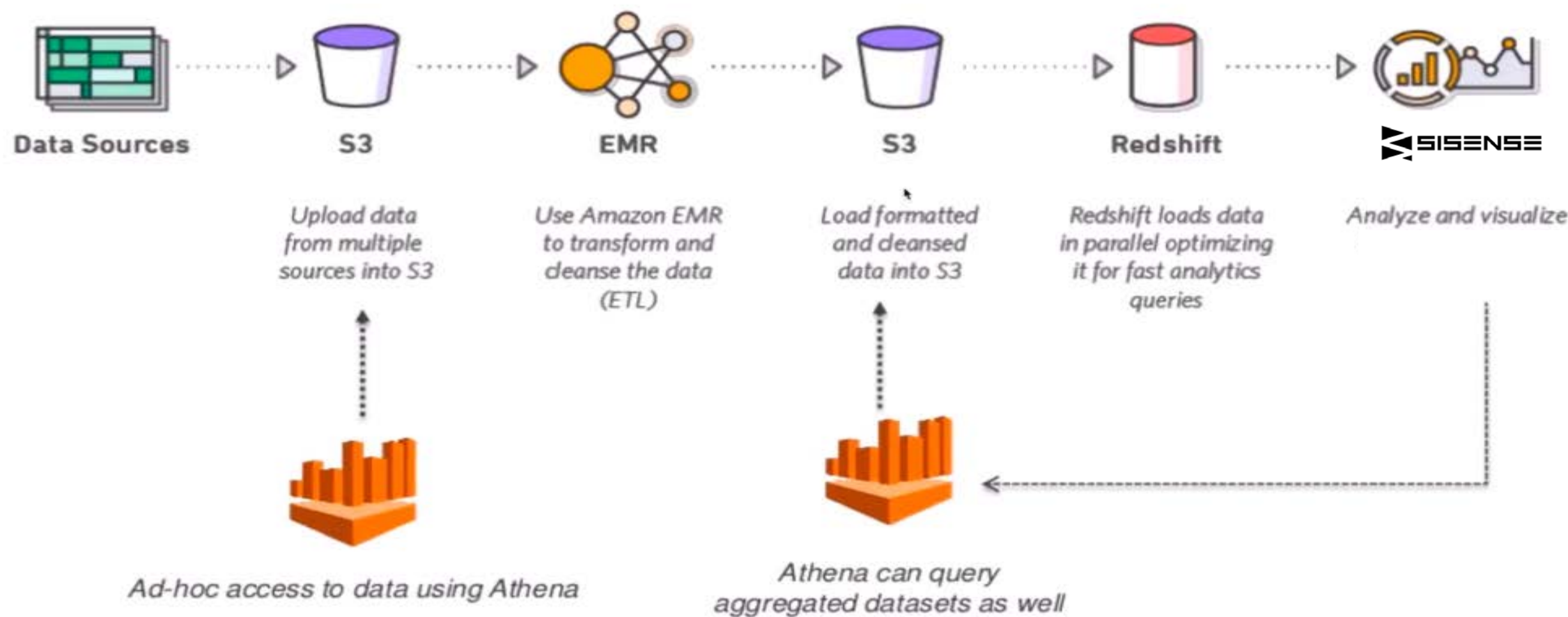
New data needs new data-flow capabilities

New

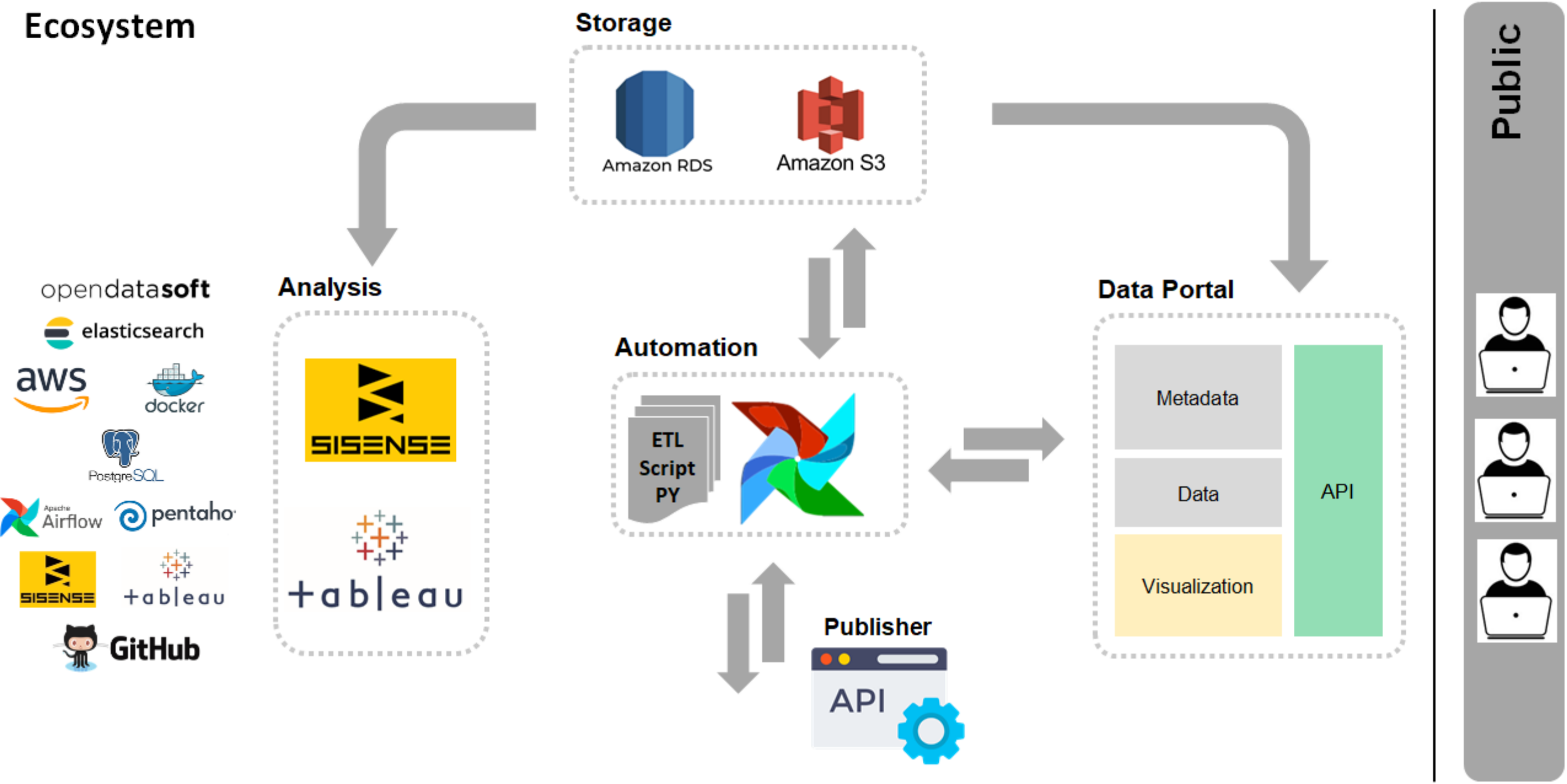


Redshift + EMR + Athena

Scale - The electricity consumption data collected once in 15 min intervals by 3 million smart meters within one year will generate 9000 TB data



Ecosystem



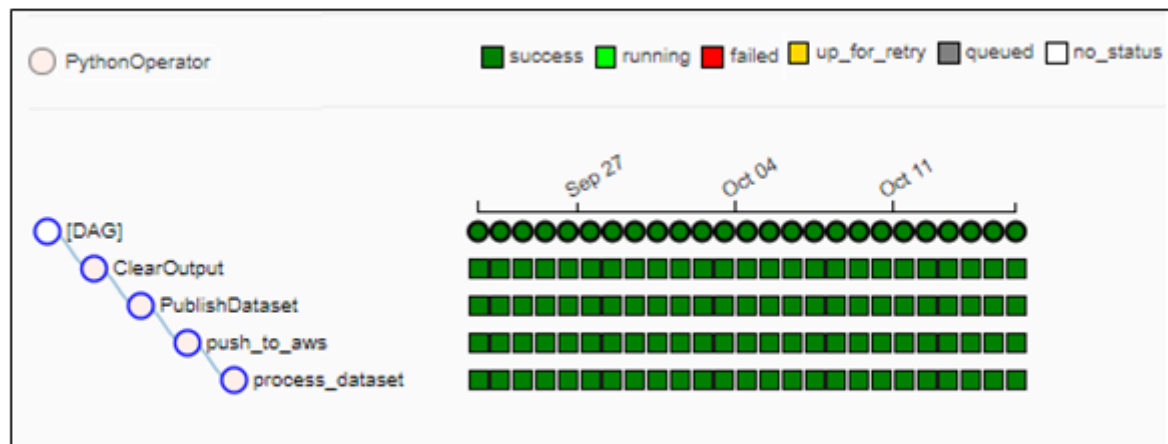
Airflow

PIPELINE EXAMPLE:

1. Process dataset (ETL)
2. Push to AWS (Storage)
3. Publish (Portal)
4. Clean files (Delete files)



Airflow						
DAGs						
Search: <input type="text"/>						
		DAG	Schedule	Owner	Recent Tasks ¹	Last Run ¹
	<input checked="" type="checkbox"/>	CME	0 3 * * *	Airflow		2020-08-23 10:13 ¹
	<input type="checkbox"/>	DWatcher	0 1 * * *	Airflow		2020-08-21 04:15 ¹
	<input checked="" type="checkbox"/>	EIA	15 9 * * *	Airflow		2020-10-16 09:15 ¹
	<input checked="" type="checkbox"/>	EndCoal	45 8 * * *	Airflow		2020-10-16 08:45 ¹
	<input checked="" type="checkbox"/>	FAO	45 4 * * *	Airflow		2020-10-17 04:45 ¹
	<input checked="" type="checkbox"/>	GlobalEnergyObservatory	0 9 1 * *	Airflow		2020-08-29 07:42 ¹
	<input checked="" type="checkbox"/>	JODI	31 5 * * *	Airflow		2020-10-17 05:31 ¹
	<input checked="" type="checkbox"/>	NOAA	5 7 * * *	Airflow		2020-10-17 07:05 ¹
	<input checked="" type="checkbox"/>	oil-price-forecasts-by-banks	30 14 20 * *	Airflow		2020-08-29 07:38 ¹
	<input type="checkbox"/>	spot-prices-for-crude-oil-and-petroleum-products	15 12 * * *	Airflow		2020-08-23 09:22 ¹
	<input checked="" type="checkbox"/>	WorldBank	45 12 * * *	Airflow		2020-10-16 12:45 ¹
	<input checked="" type="checkbox"/>	World_Coal_Plants_Database	45 5 * * *	Airflow		2020-10-17 05:45 ¹



Model Configuration

