Methodology / Process

8 categories of Tasks in the Ensights Analytics Process

Analytics as a process
(8 categories of tasks)

Acquire
Prepare
Store
Analyze
Visualize
Disseminate
Orchestrate
Operationalize

Energiewerks

https://www.energiewerks.com

Methodology / Process

8 categories of Analytics Tasks

Acquire :
Acquire raw datasets from public, proprietary and internal data stores in raw format.

Prepare / Blend :
Parse, cleanse, enhance and combine multiple datasets and reference data together in to data that can be analyzed.

Store :
Store data using various storage engines based on relational, NO SQL and in memory data stores in an efficient manner for analytics use.

Analyze:
Extract stored data and analyze using quantitative, statistical , machine learning and deep learning methodologies

Visualize:
Visualize the results in creative ways so that inferences from large amounts of data can be visualized and analyzed in a concise manner.

Disseminate:
Distribute or syndicate data, tables, narratives and images to blogs, mobile apps, enterprise apps, business intelligence tools and databases

Orchestrate:
Define a sequence comprising one or more of the tasks in a workflow to run in sequence, parallel or conditionally.

Operationalize:
Run workflows on triggers that include schedule, event, sufficiency conditions and manual stimulus

Analytics as a process
(8 categories of tasks)

Acquire
Prepare
Store
Analyze
Visualize
Disseminate
Orchestrate
Operationalize

Category One : Acquire : Data Acquisition Tasks

Data used in reporting, predictive analytics and machine learning can be obtained from public and government organizations, proprietary vendor databases and internal systems.

This data is available in various formats. Data may be in the form of a Comma Separated Value (csv) file. It may be obtained by scraping a web page for pieces of information. Crawlers will extract a hierarchy of related web pages that contain relevant information. Some proprietary data vendors have anonymous or secure ftp websites from where data can be extracted. Other proprietary data vendors may offer programmatic access to a REST application programming interface to retrieve data.

Data obtained from such services may be of different types – plain text, xml , html or json. Each of these files could be free form or conforming to a schema. Some other sources of data could be real time streaming sources of information.

For analytics purposes, all this data has to captured in real time or periodically. The data retrieved has to be stored in the raw form.

Energiewerks Ensights provides a set of tools, patterns and idioms to retrieve data from all of these data sources. Scripts to retrieve data from often used sources are available in the toolkit.

These scripts are incorporated into analytic solutions to accomplish data acquisition tasks.

Methodology / Process

Category Two : Prepare : Prepare / Blend Tasks

Data in raw form obtained from a source is not readily usable.

Data munging is a process where raw data is converted into a form that can be analyzed. The output of the data munging process is prepared datasets which are cleansed, normalized, annotated and structured data that conforms to an information model specific to the business domain.

The Energiewerks Ensights toolkit parses different formats and extracts relevant information into an efficient analytics focused data structure. The prepare process typically fills gaps in data with relevant values that may be default values or values conforming to some mathematical, statistical or numerical model.

Reference data could be added to enhance the source data. Two or more sets of data can be combined either using merge. Blended values can be calculated from the underlying raw data that may be only be obtained by combining two disparate datasets from completely different sources. This data may conform to some matching criterion with one-to-many relationships or may use an analytical model to derive values.

The software artifacts will prepare this data and store it in a memory efficient structure geared to easy analysis.

Category Three : Store :
Relational, Columnar, Graph, Document and Time Series databases.

The prepared data is further converted into a format that stores data efficiently in relational and NoSQL databases.

The flavors of databases currently supported are Relational (MySQL), Document (MongoDB), Columnar (Cassandra) and Graph (neo4j). Additional support for InfluxDB, Snowflake, Memcache and other stores planned.

Data stored in these data stores are accessed by analysts either directly from the databases if desired. Alternatively, tools exists that merge data from and to these databases in user friendly interfaces via SQL like interfaces, rest apis or micro applications developed using the dissemination toolkit.

Methodology / Process

Methodology / Process

The analyze toolkit comprises models that are use case specific. These models could be simple scripts that blend datasets. Enhancement of datasets with reference information is an additional task.

More complex tasks associated with data mining, machine learning, deep learning, statistical analysis, technical analysis and fundamental analysis scripts are built using the analyze tools.
Scripts can be written in various numerical programming packages such as python , R , Matlab and C++.

These scripts are integrated in to the analytic process via a scheduler driven , event driven or workflow driven stimulus from the Ensights toolkit.

Outputs of these models are stored in one or more databases or output as a Comma Separated Value (CSV), Excel (XLS) or Text (txt) format for further visualization or dissemination.

Category Four :
Analyze : Quantitative, Statistical, Machine and Deep Learning

Category Five : Visualize :
Concise, timely, relevant visualization of large data sets

Visual representation of source data or model outputs is generated by these the Visualize tools.

Various libraries across the python, R, Matlab , GIS (ESRI, QGIS) and Javascript stacks are used to represent data intuitively based on the use case.

Innovative techniques can be quickly incorporated in to the visualization components as they become available due to the cohesive and uncoupled nature of the scripts involved in the process.

Methodology / Process

Category Six : Disseminate :
Distribute or syndicate data, tables, narratives and images

Reports, Predictive Analytics and Research is disseminated using Email, Portable Document format (pdf) files, Excel (xls), images, tables, narratives , blogs, Slack integration and micro apps.

Additional facilities are available to syndicate content in to dashboards created within QlikSense . Pentaho and Tableau.

Other tools include online OLAP front ends, GRAPHML representations of data or geospatial data formats such as GeoJson, KML or shapefiles.

Self serve microapps for specific functional users can be developed using two rapid application development web based frameworks included in the Ensights Toolkit.

Category Seven : Orchestrate :
Concise, timely, relevant visualization of large data sets

Orchestration is the process of stitching together the above individual tasks together in to a workflow.

Various methods are available that can take care of native support on the computing platform or current customer investments. Support for Talend, Spotfire, rules based engines, Rete based fuzzy logic triggers for sufficiency based triggers are supported.

Support for notebooks such as Ipython and R notebooks is also inherently supported.

Methodology / Process

Category Eight : Operationalize : Run all cohesive tasks autonomously at enterprise scale

The operationalize toolkit consists of tools implementing three stimuli methods to trigger autonomous execution of the tools detailed above.

The first is scheduler based. Schedulers like Quartz, Tidal and Windows Task Manager are supported.

The second is event driven using AMQP compliant message brokers – Active MQ, Rabbit MQ and Kafka are supported. Any AMQP compliant proprietary messaging broker can be seamlessly integrated with appropriate configuration parameters with minimal effort.

The third paradigm uses a rest api / web service based trigger which can be used to trigger particular aspects of the analytics program either through remote applications or via an api so that externally developed applications or user facing applications can trigger the analytics processes and workflows.

Override of the autonomous process is one use case where the third paradigm can be deployed allowing the end user finer grained control without the need for IT intervention.

Fundamental Analysis Case Study

Population Weighted Degree Day Indices

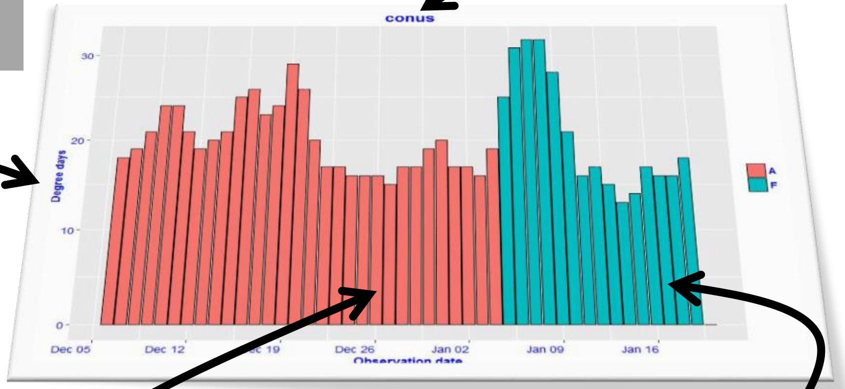
Delivery frequency : four times a day

Generated for every run of the GFS and European Ensemble. The regional and L48 Population Weighted Degree Day Indices Used for estimating natural gas Supply and Demand and price forecasting.

CENSUS DATA :

Population data obtained from the Census 2010 data set. Accounts for yearly population growth and changes based on new household creation in various parts of the country. Population granularity obtained down to individual zip codes. Grossed up by city, Metropolitan Statistical Area (MSA), individual state, EIA Region and Lower 48 states

Population weighted degree days calculated based on regional population changes obtained from Census 2010 data set.



US Lower 48
EIA Regions
Individual states
Cities
Metropolitan statistical areas (MSAs)

ACTUAL WEATHER :

Internal proprietary source used for hourly observations at 1800+ cities across the Lower 48 states. Weather observations obtained every hour on the 5's.

Easton Research Weather Index based on actual hourly weather observations for the last 30 days at 1800 + cities in the US Lower 48 states.

Easton Research Weather Index based on daily 15 day forecast at 1800+ cities in the US Lower 48 states updated 4 times a day

WEATHER FORECAST:

Internal proprietary source used for 15 day forecast. Hourly 5 day forecast is also available for 1800+ cities across the Lower 48 states.

Acquire :

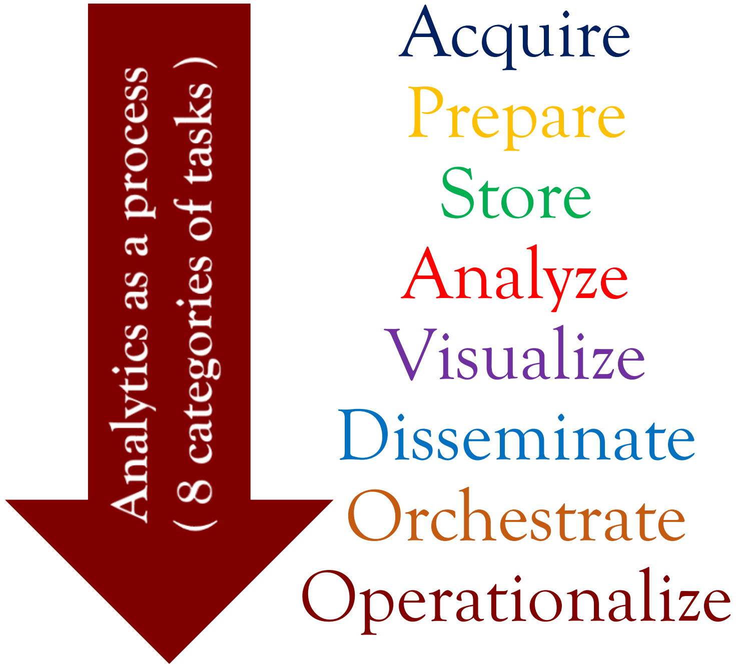
Acquire current, 0-5 day hourly and 6-15 daily forecasts(json), Census data(zip), MSA configurations (xls)

Prepare / Blend :

Parse json , zip csv, xls data, blend reference and sourced data in to facts and dimensions, aggregates

Store :

Store underlying data in raw form to file store, relation data warehouse and columnar store for analytics



Analytics as a process (8 categories of tasks)

Acquire
Prepare
Store
Analyze
Visualize
Disseminate
Orchestrate
Operationalize

Analyze:

Calculate population weights, index generation using metadata.

Visualize:

Combine current and forecast indices into histograms, tabular content and narratives that are used in dissemination.

Disseminate:

Disseminate to blogs, Slack , mobile phone apps, Tableau, Looker, QlikSense and other analytics for Supply and demand.

Orchestrate:

Orchestrate the execution of individual tasks based on schedule. Support for Task Manager, Cron, Tidal and Quartz

Operationalize:

Run workflow tasks on schedule. Report execution status

Fundamental Analysis Case Study

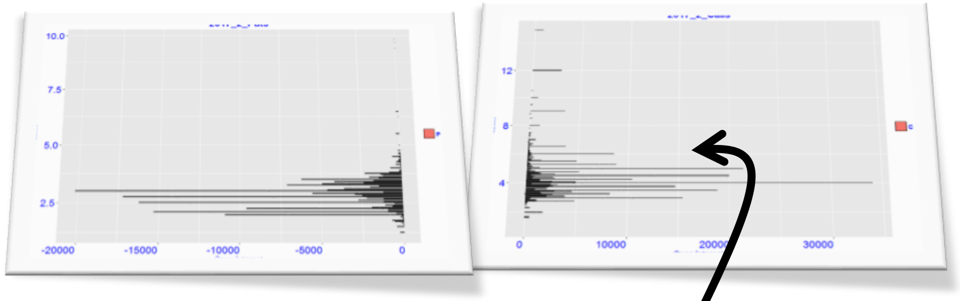
Tasks for Degree Day Indices

Top 30 contracts by volume and open interest:

Visualization of total volume and open interest on all three exchanges (Nymex, Ice and Nasdaq) where natural gas options are traded sorted in descending order with contract and strike.

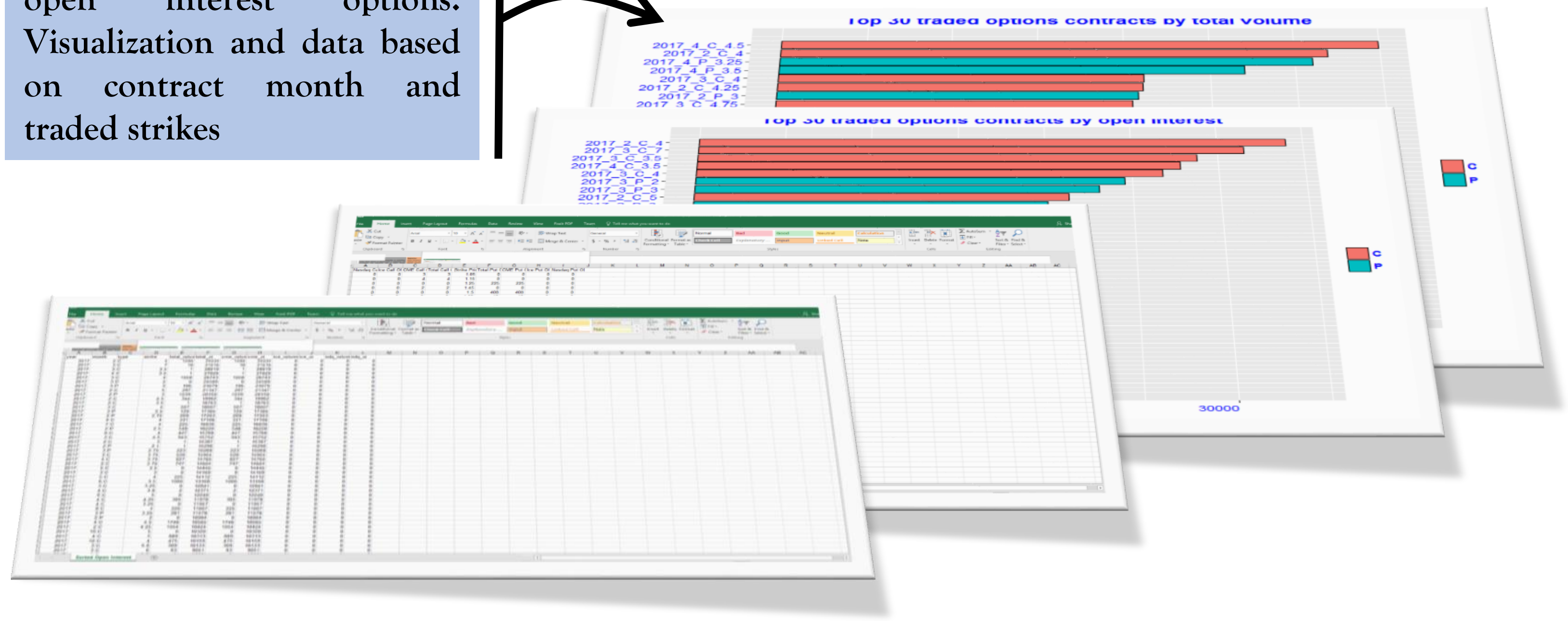
Technical Analysis Case Study

Open Interest, Volumes, Implied Volatility, Greeks



Open Interest at each strike for puts and calls by expiry month. Reports and Visualization for 12 months from prompt month

Top 30 volume traded and open interest options. Visualization and data based on contract month and traded strikes



Implied Volatility, Skew, Greeks and option structures :

Calculated Implied Volatility based on previous days settles. On demand generation of implied volatility based on intra day traded options. Volatility skew for various expiry dates. Implied volatility and skew calculations carried out using proprietary pricing library build on Quantlib. Greeks calculated using proprietary pricing library. Cash and Position Greeks calculated. Option price and greeks calculated for super set of viable ranges of underlying prices, strikes , volatility , interest rates and time to expiry. Point and click creation of option structures and associated calculation of premium, cash and position greeks.

Tasks for Open Interest, Volumes

Technical Analysis Case Study

Analytics as a process
(8 categories of tasks)

Acquire
Prepare
Store
Analyze
Visualize
Disseminate
Orchestrate
Operationalize

Acquire :
Acquire open interest and volume reports from CME, ICE and Nasdaq

Prepare / Blend :
Parse xml, csv and pdf source file. Prepare facts and dimensions based on type strikes and underlying contracts

Store :
Store file store, relation data warehouse and columnar store for analytics

Analyze:
Analyze data sets per day per contract volume, sort highest to lowest by open interest and volume and open interest.

Visualize:
Visualize puts and calls , oi and volume in most relevant formats to the trader.

Disseminate:
Disseminate to blogs, spreadsheets, Business Intelligence tools.

Orchestrate:
Orchestrate some autonomous and some manual triggers to execute workflow.

Operationalize:
Run workflow tasks on schedule and some on manual stimulus. Report execution status