# Data Warehouse Modelling Workgroup

| | |
|---|---|
| **CASE STUDY** | ALASKA & ARKANSAS - OIL, GAS, WATER PRODUCTION |
| **PROFESSOR** | JOSÉ CURTO DÍAZ (jcurto@faculty.ie.edu) |
| **TEAM MEMBERS** | TEAM A, *Ckalib N., Kathleen Co, Sarang Z., Abdulaziz A.i, Michael W., Nabil M.* |
| **DUE DATE** | 29-FEB-2020, 11:59 PM CET |

## I. DATASET ANALYSIS

The team was presented with two Enigma datasets, specifically the historical Oil, Gas, and Water production for Alaska and Arkansas as of February 2015. For the purpose of dataset analysis in this report, we discuss the following:

1. **Physical Analysis**: Covering a topical review of the raw tables, columns, data volume and data quality
2. **Logical Analysis**: Covering a brief review of the data contents and the behaviours of the same
3. **Business Analysis**: Covering the the business context of the datasets
4. **Transactional Observations**: Covering our general observations of both datasets

### A. PHYSICAL ANALYSIS

As mentioned earlier, two datasets were provided for analysis, henceforth to be called the "Alaska dataset" and the "Arkansas dataset". There are a few key differences between the two datasets. Whereas the Alaska dataset has roughly 896K rows, the Arkansas dataset has roughly 1.9M rows. The Alaska dataset has a record of oil, gas, and water production between December 1919 to October 2013. On the other hand, the Arkansas dataset has a record of oil, gas, and water production between June 1900 and December 2014, a slightly longer period.

Both the Alaska and Arkansas Datasets were setup with the same table structures carrying an API Number, Month of Production, Oil Production Volume, Gas Production Volume, Water Production Volume, and a unique serial ID per dataset. We provide further details on the table structure and data types in Table II.A.1.

**TABLE I.A.1 – RAW DATASET – COMMON TABLE DEFINITION AND DATA TYPE**

| COLUMN NAME | DEFINITION | DATA TYPE |
|---|---|---|
| API_NO | American Petroleum Institute – Details in Table II.A.2. | Fixed length integer, 14-digits, Not NULL |
| MONTH | Production Month, default set to first day of each month | Date, MM/DD/YYYY, Not NULL |
| OIL | Production volume of oil in barrels (bbls) | Float / Double |
| GAS | Production volume of gas in thousands of cubic feet (Mcf) | Float / Double |
| WATER | Production volume of water in barrels (bbls) | Float / Double |
| SERIALID | Serial ID | Integer, Not NULL |

The API Number is composed of 5 main sub-components: State, County, Unique Well Identifier, Directional Sidetrack Code, and Event Sequence Code. The first 3 parts of the code will uniquely identify a well, however the appending of the latter 2 parts of the code indicates that the code is non-persistent, i.e. subject to changes in the number of directional sidetrack and events performed onto a specific well. We provide further details on the API Number in Table II.A.2.

**TABLE I.A.2 – API NUMBER DESCRIPTION**

The API number is composed of 14 digits composed of 5 main sections: **AB-CDE-FGHIJ-KL-MN**

| API SECTION | DEFINITION | DESCRIPTION |
|---|---|---|
| AB | State Code (SOWLA Website) + 52-54 Future States + 55-61 Pseudo States. The list of SOWLA state codes are available in the Appendix. | Range 01 to 61, Integer, not NULL |
| CDE | County Code representing a county within AB State Code | Range 001-999, Integer, not NULL |

| | | |
|---|---|---|
| FGHIJ | Unique Well Identifier representing a unique well within AB State Code and CDE County Code; The unique well ID cannot be repeated within a specific AB+CDE Combination. | Range 00001 to 99999, Integer, not NULL |
| KL | Directional Side-track Code representing the number of instances of side-tracking, or if there has been an instance of drilling around a broken drill pipe or cases. For example, a "03" means that it was the third side-tracked well in the state where it is located. | Range 00 to 99, Integer, not NULL |
| MN | Event Sequence Code represents the number of operations performed on a single bore hole | Range 00 to 99, Integer, not NULL |

This unique number is the official successor to D12A number (introduced in 1966) of the American Petroleum Institute (API). It is a foundation for the management and exchange of all information about the petroleum wells in USA . The design of this Standard and the recommendations for creating the API numbers are guided by the below principles in hierarchal order:

a) Every Well and Wellbore is identified by an API Number

b) The API Number is a **unique** number

c) The API Number is a **permanent** number

d) The API Number from the Primary Assigning Authority supersedes any other identifier in public circulation

e) Every API Number is **related to a Well Origin**

f) The API Number conforms to this Standard[1]

We next analyse the data quality in the raw Alaska and Arkansas datasets. The team did not perceive significant data quality concerns with either dataset save for two main areas: Missing Values and Data Errors. The Arkansas dataset contained a few BLANK measurements under Gas, Water, and Oil. The team is uncertain about the significance of these empty cells as they came alongside some zero entries. Possibly, to differentiate between BLANK and ZERO, perhaps a blank represents a lack of attempt to extract whilst a zero represents an attempt to extract but with no successful extraction. However, holistic observations of the dataset points to the BLANK and ZERO being simply a data entry inconsistency representing the same scenario – zero extractions.

**TABLE I.A.3 – DATA QUALITY ASSESSMENT**

**API NUMBERS**

| OBSERVATIONS | ALASKA | ARKANSAS | COMMENTS |
|---|---|---|---|
| Invalid Entries | None | None | Checks: No Entry, Zero Value Entries, Non-Numeric Entries, Length not 14 |
| API Numbers w completely ZERO extractions | 39 Records | 2 Records | See Appendix 2 for full listing |
| # Distinct API Numbers | 5,267 | 14,787 | Arkansas vs Alaska: 2.81x |
| # Distinct Wells | 3,886 | 14,786 | Arkansas vs Alaska: 3.80x |

**DATE**

| OBSERVATIONS | ALASKA | ARKANSAS | COMMENTS |
|---|---|---|---|
| Oldest Date | Dec 1, 1919 | Jun 1, 1900 | |
| Newest Date | Oct 1, 2013 | Dec 1, 2014 | |
| # Ave Months Extraction per API Number | 170 Months | 129 Months | Difference: 41 Months |
| # Months w completely ZERO extractions | 556 Months | 320 Months | Difference: 236 Months |

**SERIAL ID**

| OBSERVATIONS | ALASKA | ARKANSAS | COMMENTS |
|---|---|---|---|
| # Serial ID | 895,717 | 1,911,642 | Matches with total # records |
| Max Serial_ID | 895,717 | 1,911,642 | Matches with total # records |

**GAS / WATER / OIL**

| OBSERVATIONS | ALASKA | ARKANSAS | COMMENTS |
|---|---|---|---|
| Gas Readings | Average - 8,073.07<br>Max - 13,588,118<br>Sum - 94,523,649,881 | Average - 105,525.48<br>Max - 2,473,913<br>Sum - 11,015,209,630 | Sum: Alaska 8.58 x |
| Oil Readings | Average – 687.62<br>Max – 699,367<br>Sum – 17,705,232,778 | Average – 19,766.55<br>Max - 592,911<br>Sum - 449,648,078 | Sum: Alaska 39.38 x |
| Water Readings | Average – 20,967.53<br>Max – 1,540,147<br>Sum – 19,501,702,295 | Average – 21,772.17<br>Max - 1,922,019,220<br>Sum - 7,086,710,238 | Sum: Alaska 2.75 x |

## B. LOGICAL ANALYSIS

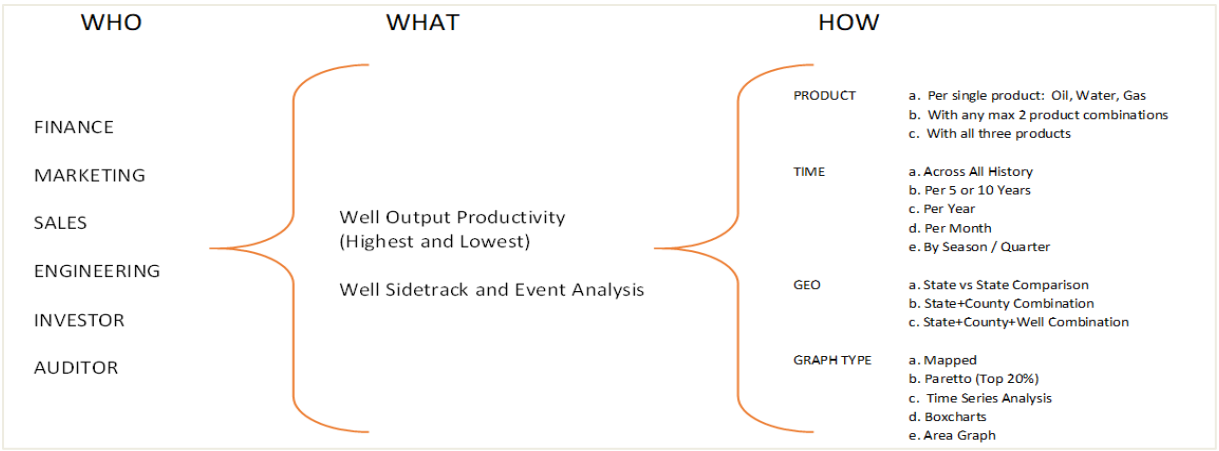
Based on the gathered information, a few interesting observations can be gathered comparing Alaska and Arkansas extractions. On average the extractions from Arkansas exceed extractions from Alaska, however the total of the extractions from Alaska far exceed those from Arkansas. Alaska vs Arkansas differences at 2.75x for Water, 8.58x for Gas, and 39.38x for Oil. Of greatest significance would be Oil which has the highest market value amongst the three resources.

It would be interesting to note as well that Arkansas started extractions 19 years earlier than Alaska and has 2-3x the number of wells compared to Alaska. However Alaska seems to be more persistent in attempting an extraction from each of its existing wells. Alaska averages 170 months extraction attempts per well, and a total of 556 months of which had completely zero extractions. In contrast, Arkansas had an average of 129 months extraction attempts per well, and a total of 320 months of which with completely zero extractions. Alaska in fact has 39 wells which have had completely zero extractions throughout history, whilst Arkansas only has 2 wells of the same.

## C. BUSINESS ANALYSIS

The Alaska and Arkansas datasets offers a wealth of information for business analysis, but the primary question we need to address will be *who* the primary target audiences are, *what* information they require, and *how* do they need to see the information. Clearly defining the business user case and specifically their business questions will enable the successful definition of a best-fit data warehouse model. The illustration below offers user persona options available:

**ILLUSTRATION I.C – BUSINESS OBJECTIVE - MAPPING OF PERSONAS**



## D. TRANSACTIONAL OBSERVATIONS

The Team discussions focused on key observations which tended towards a **DIMENSIONAL MODEL WITH STAR SCHEMA**:

1. **SMALL DATA VOLUME** – The dataset provided is small at 2.9 Million records. Even assuming an expansion from 2 to 50 states, we should still be looking at less than 500 Million records.
2. **SLOW CHANGING DIMENSIONS** – The dataset on hand is slow changing and relatively static. The information set being monitored is as well relatively static and slow to change. State and county codes in particular are completely static, whereas new well codes or changes to the same will be infrequent.
3. **PREDICTABLE, INFREQUENT TRANSACTIONS** – Transactional data comes in monthly updates. The data points in our dataset are not overwritten and each date is inserted with each production. In contrast, for example, if we decided to switch to semi-real-time monitoring enabled by IoT sensors then should be anticipating a huge surge in real-time concurrent transactions easily overwhelming a traditional model.
4. **STRUCTURED DATA** - The current data set is structured and easy to manage and organize. In contract, if we decided to expand to include site survey images or cross-reference to environmental certifications then we will find ourselves having to design a more complex model to handle unstructured data.

**ILLUSTRATION: 2015 US OIL AND GAS ACTIVITY**
Source: https://www.fractracker.org/map/national/us-oil-gas/

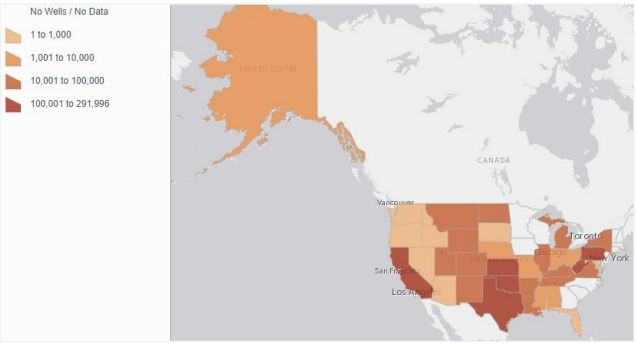
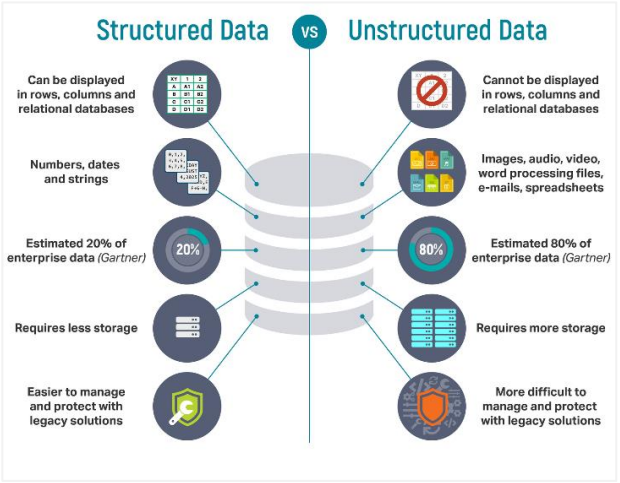



**ILLUSTRATION: STRUCTURED VS UNSTRUCTURED DATA**
Source: https://lawtomated.com/wp-content/uploads/2019/04/structuredVsUnstructuredIgneos.png



# II. Data Warehouse Approach Selection

## A. MODELLING APPROACH SELECTION – STAR SCHEMA

**DIMENSIONAL MODEL WITH STAR SCHEMA** - The dataset will be based on 3 main dimensions: Location, Date, and Operation Details and 3 measures: Oil production, Gas Production, and Water production. Because our expected users for the dataset are anticipated to come from multiple business objectives and lenses therefore we propose the creation of the below data marts:

**TABLE II.A – BUSINESS OBJECTIVE - MAPPING OF PERSONAS**

| DATA MARTS | USER SEGMENT / PURPOSE | MEASURES / VIEWS |
|---|---|---|
| (1) Resource Extraction Productivity | User Segment:<br>▪ Engineering<br>Purpose:<br>▪ To review productivity<br>▪ Identify efficiency areas<br>▪ For audit control purposes | Measures:<br>▪ Extraction Productivity<br>▪ Dis-Productivity / Inefficiencies<br>Views:<br>▪ Resources View<br>o X-Resources<br>o By Resource Type<br>▪ State-wise View<br>o X-State<br>o Arkansas Only<br>o Alaska Only<br>▪ County-wise View<br>o X-State (Default ALL)<br>o State Specific Counties |
| (2) API Information Management | User Segment:<br>▪ API Administrator<br>Purpose:<br>▪ To oversee overall well status and activity<br>▪ To govern overall well allocations by control areas | Measures:<br>▪ Well Allocations<br>▪ Well Activity Status<br>Views:<br>▪ Resources View<br>o X-Resources<br>o By Resource Type<br>▪ State-wise View<br>o X-State<br>o Arkansas Only<br>o Alaska Only<br>▪ County-wise View<br>o X-State (Default ALL)<br>o State Specific Counties |

## B. COMPARISON OF MODELLING APPROACHES

We clarify in this section the differences between each key modelling approach, and present an example of modelling under each approach.

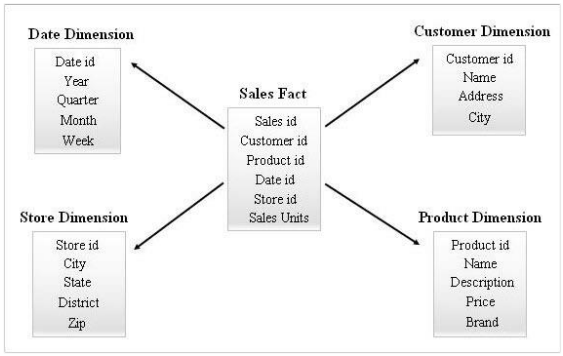
1- **STAR Schema:** The star schema consists of one or more fact tables referencing any number of dimension tables.

2- **Snowflake Schema:** The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions. "Snowflaking" is a method of normalizing the dimension tables in a star schema. When it is completely normalized along all the dimension tables, the resultant structure resembles a snowflake with the fact table in the middle.

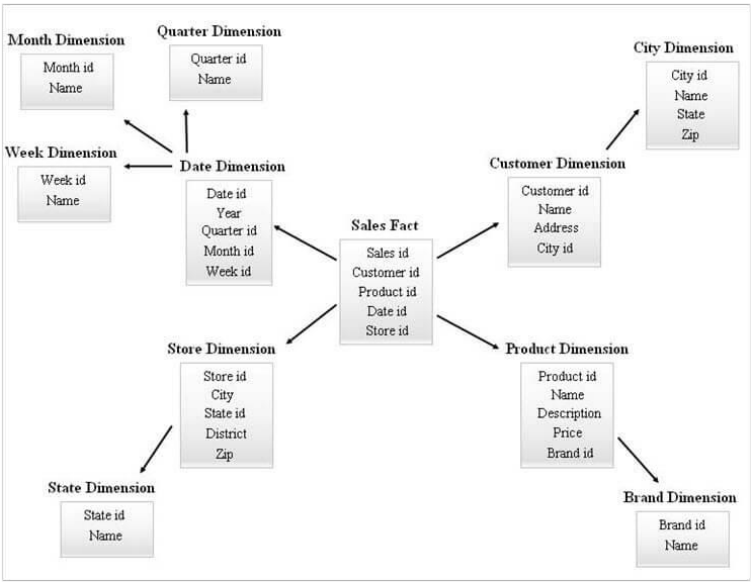
3- **Data Vault:** Data vault is a database modelling method that is designed to provide long-term historical storage of data coming in from multiple operational systems. It is also a method of looking at historical data that deals with issues such as auditing, tracing of data, loading speed and resilience to change as well as emphasizing the need to trace where all the data in the database came from. This means that every row in a data vault must be accompanied by record source and load date attributes, enabling an auditor to trace values back to the source.

**ILLUSTRATION II.B – MODELLING APPROACHES ILLUSTRATIVE EXAMPLES**

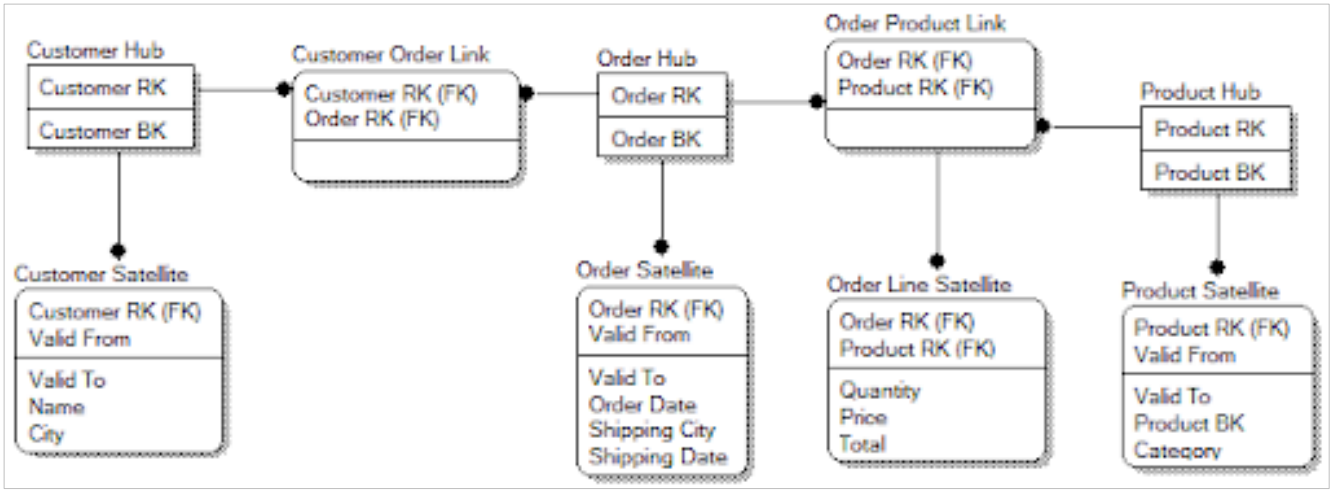
**STAR SCHEMA**




**SNOWFLAKE SCHEMA**




**DATA VAULT DESIGN**



## C. SNOWFLAKE SCHEMA & DATA VAULT INCOMPATIBILITY

Prior to determining which model would be best, we had to consider all the options available to us. Although we could have decided on a Snowflake Schema or a Data Value Model, we believed the star schema fit our use case best. We have highlighted the key points behind our reasoning below:

Here are a few more appropriate scenarios for leveraging a Snowflake Schema Model:

- Large customer dimensions where, for example, 80 percent of the fact table measurements involve anonymous visitors about whom you collect little detail, and 20 percent involve reliably registered customers about whom you collect much detailed data by tracking many dimensions
- Financial product dimensions for banks, brokerage houses, and insurance companies, because each of the individual products have a host of special attributes not shared by other products
- Multi-enterprise calendar dimensions because each organization has idiosyncratic fiscal periods, seasons, and holidays

It is our opinion that a Snowflake Schema would have been best for a much more granular dataset. Generally speaking, Ralph Kimball recommends that in most cases, Star Schemas are a better solution. Although redundancy is reduced in a normalized Snowflake, more joins are required. Kimball usually advises that it is not a good idea to expose end users to a physical Snowflake design, because it almost always compromises understandability and performance.[1]

Here are the pros and cons of a Data Vault:

- Pros
    - High Performance
    - Historical Traceability
    - Supports isolated, flexible and incremental development
- Cons
    - Data Vault requires a lot of JOIN's to derive data mart
    - Like 3NF, Data Vault is impractical for direct querying
    - De-normalization means more storage is required

Data Vault Modelling is exponentially more robust however it is more complex upfront and bigger than a Star Schema. Data Vault would be ideal to support business cases where we expect high Volume, Velocity, and Variability which is not the case for this particular business case as detailed in this report (Section I-D). A data vault may become more applicable if the business case changes and data sourcing integrates real-time / semi-real-time capture of multiple readings for instance weather readings, temperature readings, or precipitation readings, on top of their production output. Some Data Vault applicable business case may even involve triangulation of resource extractions with real-time commodity pricing.

Therefore after detailed discussion of the applicability of each modelling technique within context of this specific business case, the team concludes that STAR SCHEMA is the best approach.

# III. Data Warehouse Design

Our Star Schema Model contains one fact table entitled, "F_PRODUCTION" and 3 dimensions entitled, "DIM_LOCATION", "DIM_WELLS" and "DIM_DATE."

## A. FACT TABLES

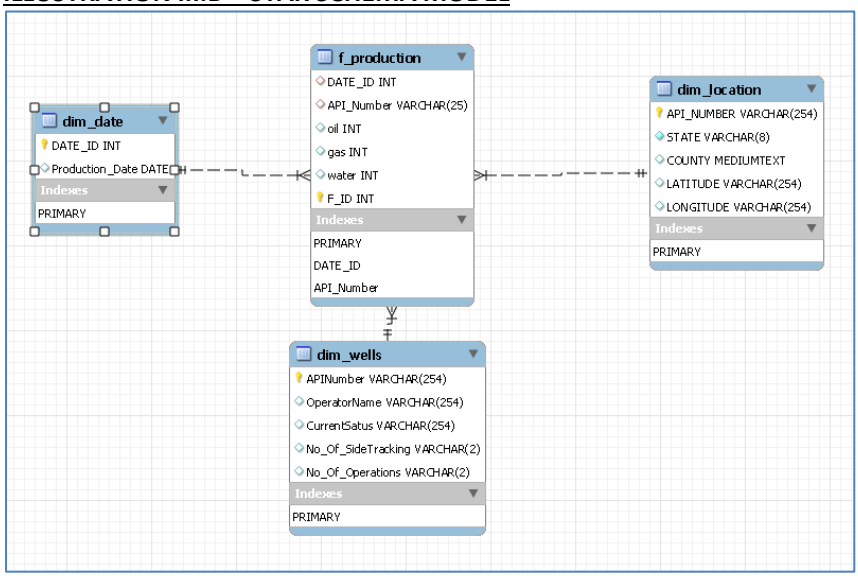
The fact table entitled, "F_PRODUCTION" consists of three main measures:

1. Oil Product
2. Gas Production
3. Water Productions.

## B. DIMENSIONS

1. **Dim_Date**: The date dimension has a key for each date, and the production date itself. Although we could have included columns such as day number, week number, or month number, we did not find it necessary
2. **Dim_Location**: The location dimension includes information regarding the geography of the wells, such as latitude, longitude, state and county
3. **Dim_Wells**: The wells dimension has the well operation information such as operator company, number, and number of operations

**ILLUSTRATION III.B – STAR SCHEMA MODEL**




## C. TABLE CORRELATIONS

- The F_PRODUCTION fact correlates with all the dimensions as showed in the indexes area as follows:
- F_PRODUCTION & DIM_LOCATION → Through a foreign key API_NUMBER
- F_PRODUCTION & DIM_DATE → Through a foreign key DATEID
- F_PRODUCTION & DIM_WELLS → Through a foreign key API_NUMBER

Each foreign key correlates with its Primary Key in each Dimension table.

# IV. APPENDIX

## API STATE AND PSEUDO-STATE NUMBER CODES

| STATE | CODE | STATE NAME | CODE | STATE NAME | CODE |
|---|---|---|---|---|---|
| Alabama | 1 | Michigan | 21 | Tennessee | 41 |
| Arizona | 2 | Minnesota | 22 | Texas | 42 |
| Arkansas | 3 | Mississippi | 23 | Utah | 43 |
| California | 4 | Missouri | 24 | Vermont | 44 |
| Colorado | 5 | Montana | 25 | Virginia | 45 |
| Connecticut | 6 | Nebraska | 26 | Washington | 46 |
| Delaware | 7 | Nevada | 27 | West Virginia | 47 |
| District of Columbia | 8 | New Hampshire | 28 | Wisconsin | 48 |
| Florida | 9 | New Jersey | 29 | Wyoming | 49 |
| Georgia | 10 | New Mexico | 30 | Alaska | 50 |
| Idaho | 11 | New York | 31 | Hawaii | 51 |
| Illinois | 12 | North Carolina | 32 | Alaska Offshore | 55 |
| Indiana | 13 | North Dakota | 33 | Pacific Coast Offshore | 56 |
| Iowa | 14 | Ohio | 34 | Northern Gulf of Mexico | 60 |
| Kansas | 15 | Oklahoma | 35 | Atlantic Coast Offshore | 61 |
| Kentucky | 16 | Oregon | 36 | | |
| Louisiana | 17 | Pennsylvania | 37 | | |
| Maine | 18 | Rhode Island | 38 | | |
| Maryland | 19 | South Carolina | 39 | | |
| Massachusetts | 20 | South Dakota | 40 | | |

## API NUMBERS HAVING COMPLETELY ZERO EXTRACTION RESULTS

**ALASKA DATASET**

| API_number | Start_Date | End_Date | No Readings | Gas | Oil | Water |
|---|---|---|---|---|---|---|
| 03103109750000 | 1/1/1990 | 11/1/2014 | 287 | 0 | 0 | 0 |
| 03139118910000 | 1/1/1992 | 1/1/2008 | 193 | 0 | 0 | 0 |

**ARKANSAS DATASET**

| API_number | Start_Date | End_Date | No Readings | Gas | Oil | Water |
|---|---|---|---|---|---|---|
| 50133101440200 | 10/1/2002 | 10/1/2013 | 131 | 0 | 0 | 0 |
| 50231200190000 | 10/1/2004 | 10/1/2013 | 109 | 0 | 0 | 0 |
| 50231200170000 | 10/1/2004 | 10/1/2013 | 109 | 0 | 0 | 0 |
| 50133205390000 | 11/1/2004 | 10/1/2013 | 108 | 0 | 0 | 0 |
| 50231200220000 | 11/1/2004 | 10/1/2013 | 108 | 0 | 0 | 0 |
| 50029227390000 | 10/1/1997 | 12/1/2005 | 99 | 0 | 0 | 0 |
| 50733200110100 | 9/1/2005 | 10/1/2013 | 98 | 0 | 0 | 0 |
| 50133100020100 | 1/1/2006 | 10/1/2013 | 94 | 0 | 0 | 0 |
| 50029214510000 | 3/1/2007 | 10/1/2013 | 80 | 0 | 0 | 0 |
| 50029231870000 | 4/1/2008 | 10/1/2013 | 67 | 0 | 0 | 0 |
| 50029231830000 | 4/1/2008 | 10/1/2013 | 67 | 0 | 0 | 0 |
| 50133205650000 | 12/1/2008 | 10/1/2013 | 59 | 0 | 0 | 0 |
| 50283200600000 | 1/1/2010 | 10/1/2013 | 46 | 0 | 0 | 0 |
| 50733200940100 | 8/1/1998 | 10/1/2013 | 40 | 0 | 0 | 0 |
| 50733203390000 | 10/1/2010 | 10/1/2013 | 37 | 0 | 0 | 0 |
| 50133204890000 | 10/1/2010 | 10/1/2013 | 37 | 0 | 0 | 0 |
| 50009200170000 | 1/1/2003 | 6/1/2005 | 29 | 0 | 0 | 0 |
| 50103203980000 | 10/1/2002 | 2/1/2005 | 29 | 0 | 0 | 0 |
| 50009200069000 | 6/1/2003 | 6/1/2005 | 25 | 0 | 0 | 0 |
| 50733202980200 | 1/1/2012 | 10/1/2013 | 22 | 0 | 0 | 0 |
| 50029226030000 | 6/1/1996 | 12/1/1997 | 19 | 0 | 0 | 0 |
| 50283201640000 | 9/1/2012 | 10/1/2013 | 14 | 0 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 50029216280000 | 10/1/1987 | 11/1/1988 | 14 | 0 | 0 | 0 |
| 50023200390000 | 1/1/2013 | 10/1/2013 | 10 | 0 | 0 | 0 |
| 50029206420000 | 3/1/1982 | 11/1/1982 | 9 | 0 | 0 | 0 |
| 50029202290000 | 10/1/1990 | 5/1/1991 | 8 | 0 | 0 | 0 |
| 50029220110200 | 12/1/2003 | 4/1/2004 | 5 | 0 | 0 | 0 |
| 50029207200000 | 8/1/1982 | 12/1/1982 | 5 | 0 | 0 | 0 |
| 50029232760000 | 1/1/2009 | 4/1/2009 | 4 | 0 | 0 | 0 |
| 50029231070000 | 4/1/2003 | 7/1/2003 | 4 | 0 | 0 | 0 |
| 50029206900000 | 8/1/1982 | 11/1/1982 | 4 | 0 | 0 | 0 |
| 50029223740000 | 1/1/2009 | 4/1/2009 | 4 | 0 | 0 | 0 |
| 50029231720000 | 10/1/2003 | 11/1/2003 | 2 | 0 | 0 | 0 |
| 50029231450000 | 5/1/2003 | 6/1/2003 | 2 | 0 | 0 | 0 |
| 50029206570000 | 8/1/1982 | 8/1/1982 | 1 | 0 | 0 | 0 |
| 50029223120100 | 3/1/2013 | 3/1/2013 | 1 | 0 | 0 | 0 |
| 50029224140000 | 1/1/2000 | 1/1/2000 | 1 | 0 | 0 | 0 |
| 50023200150000 | 9/1/2004 | 9/1/2004 | 1 | 0 | 0 | 0 |
| 50733201190000 | 3/1/1981 | 3/1/1981 | 1 | 0 | 0 | 0 |

# V. References

**1. API NUMBER GUIDELINES [ONLINE]**
Available at: https://penerdeq.ihsenergy.com/dynamic.splashscreen/documents/IHS%20API%20Numbering%20Guidelines.pdf [Accessed 10 February 2020]

**2. STAR AND SNOWFLAKE SCHEMAS. 2020. STAR AND SNOWFLAKE SCHEMAS. [ONLINE]**
Available at: https://www.oracle.com/webfolder/technetwork/tutorials/obe/db/10g/r2/owb/owb10gr2_gs/owb/lesson3/starandsnowflake.htm. [Accessed 28 February 2020].

- o -