**ADVANDB MCO2: Data Warehousing and Online Analytical Processing (OLAP)**

Ryan Austin Fernandez

CS-ST Student

2401 Taft Avenue,

Manila, Philippines 1004

(+63)916-496-5071

Alden Luc Hade

CS-ST Student

2401 Taft Avenue,

Manila, Philippines 1004

(+63)922-431-0831

ryan\_fernandez@dlsu.edu.ph

alden\_hade@dlsu.edu.ph

Clarisse Felicia Poblete

CS-ST Student

2401 Taft Avenue,

Manila, Philippines 1004

(+63)915-624-5417

clarisse\_poblete@dlsu.edu.ph

Marc Dominic San Pedro

CS-ST Student

2401 Taft Avenue,

Manila, Philippines 1004

(+63)977-817-5699

marc\_sanpedro@dlsu.edu.ph

**1. INTRODUCTION**

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce nec pellentesque risus. Vestibulum euismod, lacus et elementum ultricies, nulla ex luctus mauris, nec euismod quam ipsum sit amet tellus. Nam sit amet congue ligula. Maecenas in lectus congue, ornare orci quis, vestibulum velit. Maecenas posuere eget turpis eu semper. Nullam luctus sapien in elementum fermentum. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris aliquam metus ut justo commodo, sed malesuada justo venenatis. Maecenas lorem quam, sagittis ac tortor imperdiet, pharetra consectetur ex.

Nunc sit amet fringilla ipsum. Duis sed mattis augue, vel ultricies ex. Fusce commodo tincidunt ipsum, ac tempor diam faucibus sed. Sed vitae diam eget lectus tincidunt placerat eu sit amet velit. Cras euismod nisi mi, in maximus ex luctus in. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Fusce blandit metus nulla, non sagittis nibh ullamcorper vitae. Sed at dictum erat.

**2. DIMENSIONAL MODEL**

The schema that has been created for this project mainly deals with land and agricultural data as the primary facts. From the CBMS database, the hpq\_hh, hpq\_mem, hpq\_crop, hpq\_alp, and hpq\_arcdp\_mem tables were transformed into the group’s constellation schema, the former two holding most of the dimensions and the latter three holding most of the facts. The entire schema can be seen in Figure 1.

**2.1. Contents**

There are three fact tables in the schema: the crop table, the arcdp table, and the land\_parcel table.

The crop table contains facts regarding the crops produced by a certain household. The measure of this fact table is the crop volume crop\_vol, which will be aggregated using sum and average.

The arcdp table contains information about the Agrarian Reform Community Development Program beneficiaries. The measures of this table are the age, calculated grade, and daily work hours of the beneficiary. All values will be averaged.

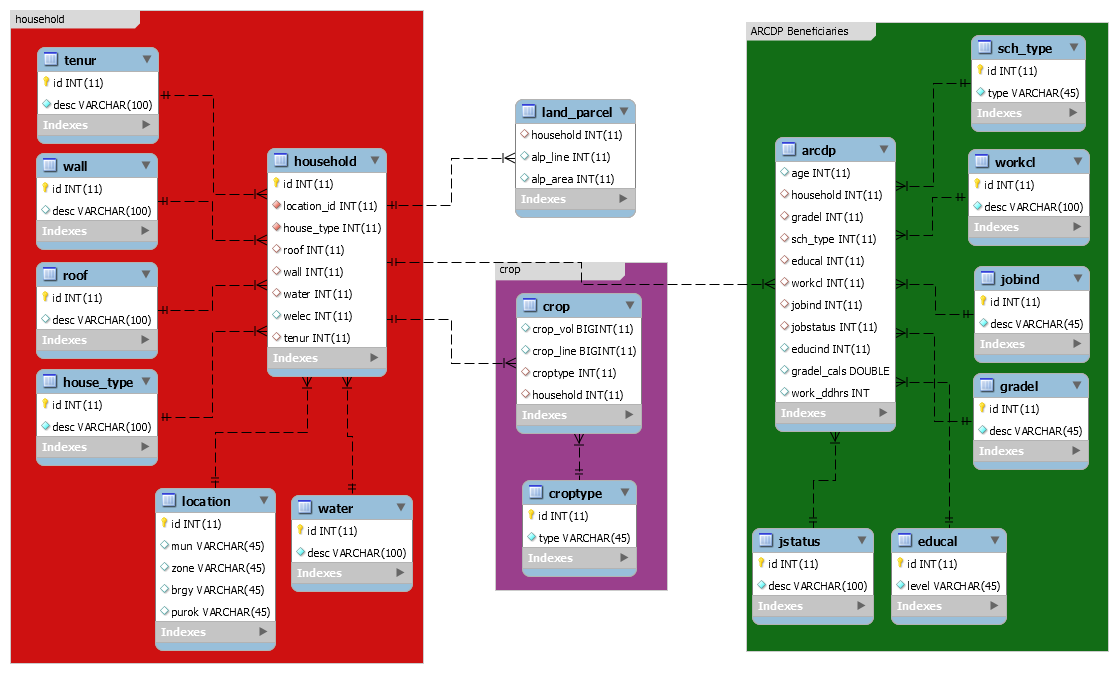
The land\_parcel table contains facts about the land owned by the households. The measure of this table is the land area, which will be aggregated using sum and average functions.

**2.2. Dimensions**

The dimensions that were used were categorical columns from the selected tables in the CBMS database. For the crop table, “croptype” is the type of crop farmed, which may be Sugarcane, palay, corn, coffee, or other.

For the households, ”house\_type” is the type of house, which may be single house, duplex, multi-unit residential, commercial/ industrial/agricultural/house, or others. “location” comprises the unique municipality, barangay, zone, and purok of the household. “wall” refers to the type of wall and “roof” is the type of roof of a household, both of which may be strong materials, light materials, salvaged materials, mixed but predominantly strong materials, mixed but predominantly light materials, mixed but predominantly salvaged materials, or not applicable. “tenur” is the tenure of the house owner, which can be owner of lot; renting lot; owner of house, renting lot; owner of house, not renting lot with consent of owner; owner of house, not renting lot without consent of owner; rent-free house and lot with consent; rent-free house and lot without consent; living in public space with rent; living in public space without rent; and other. “water” is the household’s source of water which may be own faucet, shared faucet, own deep well, shared deep well, shallow well, dug well, protected spring, unprotected spring, natural, peddler, bottled water, or others.

For the beneficiaries, “jobind” is whether the arcdp beneficiary is employed or unemployed., “jstatus” is the current status of the beneficiary at their job, which may be permanent, short-term, or working on different jobs week to week. “sch\_type” is whether the beneficiary is at a public or private school. “workcl” is what kind of employee the beneficiary is which may be for private household, for private business, for government, self-employed with no employee, employer in family business, working with pay on family-owned business, or working without pay on family-owned business. “educal” is the highest educational attainment and “gradel” is the current grade which may be none, day care, preschool, grades 1 to 12, 1st to 3rd year PSPS/N-T/TV, 1st to 4th



**Figure 1 – The Constellation Schema**

year college, post grad, ALS elementary or secondary, SPED elementary or secondary, Grade school, High School, Post secondary, college, or MD/PhD.

To distinguish which columns to use and under what context, a rudimentary understanding of statistics was used. For each table, having joined the hpq\_hh and hpq\_mem tables with the hpq\_crop, hpq\_alp, and hpq\_arcdp\_mem tables, the four classifications of variables: nominal, ordinal, interval, and ratio (Types of Variable, n.d.); were used to distinguish between dimension and fact column.

Nominal variables, which may be numeric or non-numeric, represent categories but have no intrinsic order. In the constellation schema, these were translated into the dimension tables croptype, house\_type, jobind, jstatus, location, roof, sch\_type, tenur, wall, water, and workcl. Ordinal variables also represent categories, but have an intrinsic order, such as the dimensions educal, and gradel. Since it would make no sense to perform aggregate functions on these types of variables since they hold no meaningful intrinsic value, it is better to classify the more meaningful data according to these types of variables, which means they are more suitable as dimensions.

Interval and Ratio variables do have a meaningful value, their only difference being that a zero in a ratio variables means that there is no quantity represented, a true zero. As such, there is meaning in adding and subtracting these values, which would make them ideal for aggregate functions such as sum and average. These variables ended up being the measures in the fact table such as crop\_vol, alp\_area, age, gradel\_calc, and workddhrs.

**2.3. Schema and Issues**

A constellation schema was implemented, with the primary fact tables being land\_parcel, crop, and arcdp. The crop table references a croptype table. All three fact tables reference household which itself references the tenur, wall, roof, house\_type, location, and water tables. The arcdp table also references the sch\_type, workcl, jobind, gradel, educal, and jstatus tables.

A constellation was selected to allow for more than one analytical view for data. Since the original CBMS database had three tables pertaining to agricultural information, a constellation was a prime opportunity to extract as much information as possible.

The initial issue with the design was that, ideally, a data warehouse would have fact tables referencing other dimension tables that have more data about that particular dimension. For example, a sales table bay have a branch dimension, but the branch table would have additional location and manager attributes. The original CBMS database did have a semblance of a star schema but , the most common dimension tables referenced were only two, although three were required for the MCO. After consultation, it was said that extracting individual data columns from the tables could account for more dimensions, so the aforementioned columns in section 2.2. were extracted into tables. The next problem is the lack of additional data in these tables. Unlike the branch table, the only other data in these tables is a descriptor, which would limit the possibilities of future slice and dice operations. Since the original CBMS database had no further data, this was simply accepted as a limitation.

**3. ETL PROCESS**

**3.1. Code Description**

The team decided to create their own ETL tool using MySQL. The ETL process was done using a Java extraction program and a MySQL transformation and loading script. The Java program read from the CSV files, loading them into the original CBMS schema. The resultant schema was then queried and manipulated to transform its data into the new schema. The results of the queries were loaded into the new schema using INSERT INTO… SELECT statements.

**3.2. Extraction**

The original data source were seventeen comma-separated values files, each file representing one relation. The Java extraction tool coded by one of the team members simply extracted the table name and all the column names to be used in the INSERT INTO <table name> (<col 1>[,<col2>...]). It then read each row to be added to the VALUES <tuple1>... part. After this, it established a connection with the database and executed the INSERT statements. This program was run for each csv file until the entire schema was loaded.

A data dictionary, which listed each table and the columns, along with a column description,  was also provided. This was used to generate the original empty schema.

The resultant database was queried in the next step, which constitutes the extraction step.

**3.3. Transformation**

For most of the dimension tables, the data dictionary was consulted in populating the tables with their actual meaning.

For the location, however, each location is a unique combination of municipality, barangay, zone, and purok. Each combination was given a unique id by giving the new table an autoincrement primary key and executing the following script.

INSERT INTO location(mun, zone, brgy, purok)

SELECT distinct mun, zone, brgy, purok FROM hpq\_hh;

For the household table, the location fields had to be transformed into the new ones generated by the previous query, so joining with the location table was required. The script is as follows.

INSERT INTO household

SELECT H.id, L.id AS location, house\_type,roof,wall,water,welec,tenur

FROM db\_hpq.hpq\_hh H INNER JOIN location L

ON H.mun = L.mun AND H.zone = L.zone AND H.brgy = L.brgy AND H.purok = L.purok

GROUP BY H.id;

For the crop and land\_parcel tables, they simply had to be loaded into the new database after projecting off some unnecessary columns.

The query for crop is

INSERT INTO crop(crop\_vol,crop\_line,croptype,household)

SELECT crop\_vol, crop\_line, croptype, hpq\_hh\_id AS household

FROM db\_hpq.hpq\_crop C INNER JOIN household H ON C.hpq\_hh\_id = H.id;

The query for land\_parcel is

INSERT INTO land\_parcel

SELECT id AS household, alp\_line, alp\_area

FROM db\_hpq.hpq\_alp A INNER JOIN household H

ON a.hpq\_hh\_id = H.id;

For the arcdp table, each beneficiary had to be associated with the community member, so the hpq\_arcdp\_mem table had to be joined with the hpq\_mem table. After the join, only a select few columns had to be retained. The query is as follows:

INSERT INTO arcdp

SELECT age, household, gradel, sch\_type, educal, workcl, jobind, jstatus, educind, gradel\_calc, work\_ddhrs

FROM (SELECT A.hpq\_hh\_id AS household, arcdp\_mem\_refno as memno

FROM db\_hpq.hpq\_arcdp\_mem A INNER JOIN household H

ON H.id = A.hpq\_hh\_id) X

INNER JOIN db\_hpq.hpq\_mem M

ON M.memno = X.memno AND M.id = X.household;

**3.4. Loading**

The new schema had primary keys and foreign key, which would enforce key integrity and referential integrity. During the loading of the three fact tables, MySQL complained about foreign key constraint violation. This would only be because the hpq\_crop, hpq\_arcdp\_mem, and hpq\_alp tables had rows that were not connected to any row in hpq\_hh. Since nothing could be done about the missing data in hpq\_hh, an INNER JOIN between the table in question and the hpq\_hh table was performed, eliminating the rows in the three peripheral tables that were not referencing an existent household.

**4. OLAP QUERIES**

**4.1. Sample Queries**

*4.1.1. Roll-up*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce nec pellentesque risus. Vestibulum euismod, lacus et elementum ultricies, nulla ex luctus mauris, nec euismod quam ipsum sit amet tellus. Nam sit amet congue ligula. Maecenas in lectus congue, ornare orci quis, vestibulum velit. Maecenas posuere eget turpis eu semper. Nullam luctus sapien in elementum fermentum. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris aliquam metus ut justo commodo, sed malesuada justo venenatis. Maecenas lorem quam, sagittis ac tortor imperdiet, pharetra consectetur ex.

*4.1.2. Drill-down*

Nunc sit amet fringilla ipsum. Duis sed mattis augue, vel ultricies ex. Fusce commodo tincidunt ipsum, ac tempor diam faucibus sed. Sed vitae diam eget lectus tincidunt placerat eu sit amet velit. Cras euismod nisi mi, in maximus ex luctus in. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Fusce blandit metus nulla, non sagittis nibh ullamcorper vitae. Sed at dictum erat.

*4.1.3. Dice and Slice*

Aenean quis velit lorem. Maecenas vestibulum, magna quis gravida fermentum, ante sapien elementum est, eu convallis dui sem in justo. Proin suscipit aliquam tortor, mollis dignissim elit mollis sed. Pellentesque laoreet ligula at dignissim pulvinar. Integer at magna mi. Vestibulum mi tortor, vestibulum in lacus eu, lobortis sodales sem. In vitae euismod magna, non tincidunt augue. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Donec interdum laoreet nisi id rhoncus.

**4.2. Queries**

*4.2.1. Crop Analytics*

Aenean quis velit lorem. Maecenas vestibulum, magna quis gravida fermentum, ante sapien elementum est, eu convallis dui sem in justo. Proin suscipit aliquam tortor, mollis dignissim elit mollis sed. Pellentesque laoreet ligula at dignissim pulvinar. Integer at magna mi. Vestibulum mi tortor, vestibulum in lacus eu, lobortis sodales sem. In vitae euismod magna, non tincidunt augue. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Donec interdum laoreet nisi id rhoncus.

*4.2.2. Land Parcel Analytics*

Mauris faucibus velit a urna commodo commodo. Cras ligula nisi, porttitor ultricies ante eget, finibus condimentum augue. Quisque finibus arcu sit amet tempus faucibus. Cras eleifend, turpis vel aliquam cursus, leo nulla euismod ex, quis rutrum nulla mauris eget quam. Ut finibus ultrices risus, mollis vehicula elit tempus eget. Aenean cursus, diam vel venenatis gravida, purus ligula porttitor mauris, in vulputate felis sem sed turpis. Mauris posuere tellus non sapien commodo viverra. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nunc molestie metus in dui vulputate eleifend. Etiam accumsan aliquet felis. Curabitur id metus mauris. Ut posuere tortor nisi, id mollis enim aliquam vitae. Sed et vestibulum nisl. Aenean tempor risus eu convallis commodo. Vestibulum bibendum vestibulum convallis. Fusce ultrices neque et velit porta, ac fermentum quam consectetur.

*4.2.3. ARCDP Analytics*

Cras porttitor augue ut porttitor congue. Vestibulum vestibulum, justo a placerat condimentum, turpis massa dictum massa, vel bibendum libero mi nec eros. Integer auctor, ante sed vulputate elementum, sapien nisi faucibus arcu, non feugiat leo mi eu quam. Donec consectetur risus vitae nisi eleifend finibus. Curabitur luctus neque eu sem gravida commodo. Sed lacus justo, lobortis a odio sed, tempus vestibulum ex. Suspendisse iaculis, velit sit amet scelerisque sagittis, sem metus varius leo, sit amet volutpat sem enim a nisl. Nam luctus lectus nec dolor pretium, eu rhoncus quam efficitur.

**5. RESULTS AND ANALYSIS**

Quisque lectus eros, varius non dictum in, semper et neque. Fusce sed convallis sem, eget ultrices eros. Nullam consequat lacus tellus, nec congue diam consectetur porta. Praesent vulputate ligula eu enim feugiat, quis malesuada ipsum rutrum. Quisque eu vehicula neque. Curabitur feugiat mi sollicitudin dui faucibus, et tempor metus varius. Maecenas ut tempor erat. Mauris quis justo fringilla, lobortis nibh quis, euismod lacus. Quisque felis magna, scelerisque at magna sed, aliquam ultricies ex. Sed sapien arcu, rutrum nec efficitur nec, volutpat non arcu. Integer tempor mi ut enim sodales, vel elementum nisl hendrerit. Aenean at cursus velit, vestibulum fringilla dui. Integer vel pharetra est. Proin dui ante, pulvinar rhoncus felis sed, sollicitudin hendrerit sem. Curabitur venenatis enim efficitur, porta neque at, efficitur nulla. Donec interdum odio eget semper tincidunt.

**6. CONCLUSION**

Aliquam in risus tincidunt, gravida risus in, laoreet sapien. Vivamus ac velit nulla. Sed ligula magna, volutpat at libero nec, blandit luctus lorem. In a est consequat ex bibendum gravida mollis vitae metus. Morbi euismod rutrum fringilla. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin rutrum neque tellus, sed luctus risus pretium at. Mauris porta ac massa vitae vulputate. Nam ultricies sed ex eget tincidunt. Donec lobortis tristique massa ac ultricies.

Mauris at orci ut justo consectetur ultrices. Cras a augue consequat, dignissim leo non, rhoncus arcu. Integer fermentum dolor vitae metus pretium, quis euismod dui posuere. Maecenas blandit iaculis lacus in laoreet. Nulla porta diam sit amet sem mattis faucibus. Morbi quis enim vestibulum tortor egestas maximus ut at nulla. Vestibulum maximus pretium neque, quis finibus magna tincidunt in. Ut vel consectetur est, non eleifend turpis. Suspendisse massa nulla, ultricies et suscipit id, pellentesque at tellus.

**7. REFERENCES**

1. Silberschatz, A., Korth, H. & Sudarshan, S. (2010). Database System Concepts, 6th Edition, McGraw-Hill Book Co. Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.