**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**

In 2013, the Angelo King Institute of Economic Business Studies partnered with the College of Computer Studies to develop a digital data collection, called the **Community-Based Management System** (CBMS). The data was collected digitally through mobile devices and a web portal. The dataset spans a number of different regions in the Philippines; however, only the Marinduque and Palawan CBMS datasets were used in this Major Course Output (MCO).

For this project, the group used the CBMS dataset and extracted select data from it into a new model, which was based from a schema designed to easily accommodate online analytical processing (OLAP). The group also developed a simple web application that allows the user to extract data from this new model, such as the bio-profile of different households and individual household members, their educational attainment and work experience, their social indicators, and criminal or mortuary events. It goes into a very deep level of detail in some respects, and depending on the parameters fed into the web application, even more detail can be derived about Palawan and/or Marinduque residents from this data.

The goal of this project is for students to realize the use of data warehousing, and the importance of online analytical processing, as well as how the two are related. Also, via this project, we will realize the correlation between good database design and snappier performance.

The data model used in this project is capable of supporting the needs of a number of different cooperative federations and non-governmental organizations, namely:

* 1. **Magsasaka at Siyentipiko para sa Pag-unlad ng Agrikultura (MASIPAG)**

This is a farmer-led network of  NGOs working towards the sustainable use and management of biodiversity through farmers’ control of genetic and biological resources, agricultural production and associated knowledge.

* 1. **Philippine Rural Reconstruction Movement (PRRM)**

The PRRM is a non-governmental organization formed in 1952, based on rural reconstruction experiences in China in the early 1900s, with the goal of assisting the poorer members of society in the Philippines.

* 1. **Labor Education and Research Network (LEARN)**

LEARN is a non-governmental organization that provides various services primarily to workers, whether informal or formal in the nature of their labor. Founded in October 1986, its core programs are education, research, publications, solidarity, and networking.

* 1. **National Cooperation of Cooperatives (NCCO)**

This organization is the largest cooperative federation in the Philippines, with 760 member cooperatives and non-governmental organizations spanning 77 provinces and 130 cities and municipalities as of 2015. Among these cooperatives are the Marinduque Social Action Multipurpose Cooperative, and the SRT Narra Cooperative of Palawan.

It allows these organizations to formulate appropriate intervention programs by providing appropriate data needed for planning. Given a certain CBMS dataset that contains data for more provinces and municipalities, these organizations can more easily devise plans for more locales and provide services to a wider variety of locations.

**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, and daily work hours of the beneficiary. All values will be averaged.

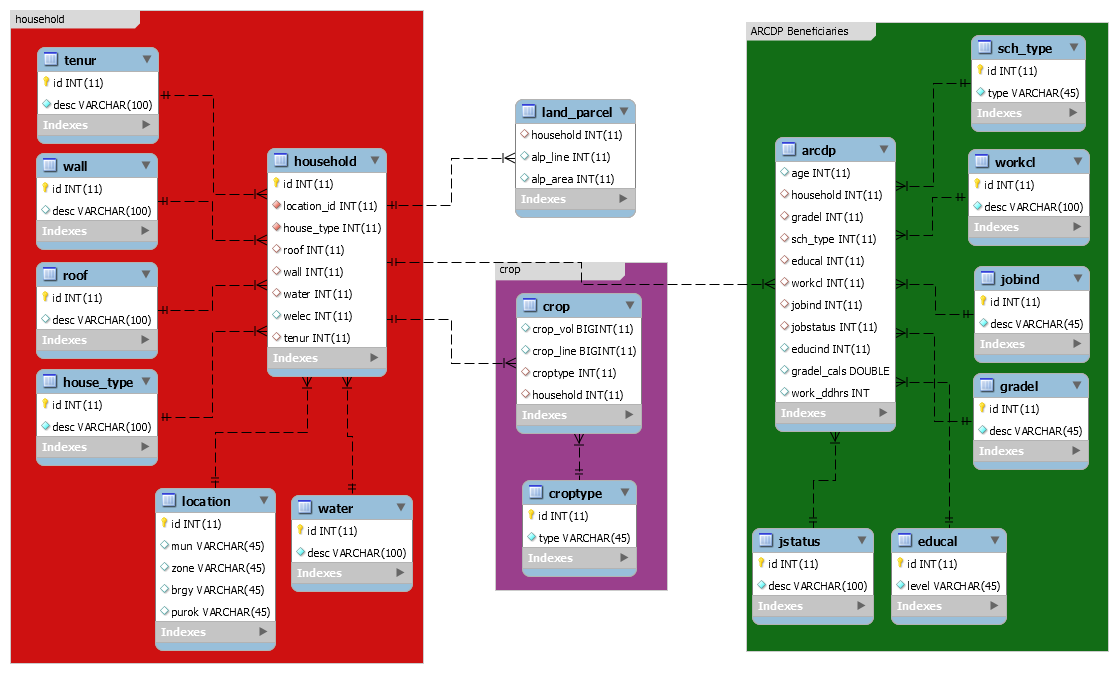
The land\_parcel table contains facts about the land owned by the households. The measure of this table is the land area, which will be aggregated using sum and average functions.

**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, 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 LEFT 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**

The main purpose of the application is threefold: to determine the typical qualities of residences and the correlation of this to the size of their land and to determine the correlation of the same qualities and crop type to the volume of their crops, and to determine the typical residence, work, and educational qualities of the ARCDP beneficiaries in relation to their age, GPA, and total daily work hours.

To query the data warehouse, a QueryBuilder class was created, which queries the constellation schema based on the dimensions selected for drill down/rollup and slice and dice.

**4.1. Sample Queries**

For the purposes of discussion, assume the base OLAP query is viewing the average age, grade, and work hours of all ARCDP beneficiaries classified by highest educational attainment. The query for this as generated by QueryBuilder is

SELECT jobind.desc AS jobind ,educal.desc AS educal ,AVG(age) AS avgAge,AVG(work\_ddhrs) AS avgWorkHrs

FROM (SELECT \* FROM ARCDP A INNER JOIN household H ON A.household = H.id) X LEFT JOIN jobind ON X.jobind = jobind.id LEFT JOIN educal ON X.educal = educal.id

GROUP BY jobind,educal

ORDER BY avgAge DESC,avgWorkHrs DESC

Which yields 41 rows, the first 10 of which are in Table 1.

**Table 1 – ARCDP Beneficiaries Classified by Job Index and Highest Educational Attainment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Job Index** | **Highest Educational Attainment** | **Average Age** | **Average Work Hours** |
| Employed | Grade 5 | 69.0000 | 8.0000 |
| Employed | 4th year College or higher | 63.0000 | 16.0000 |
| Employed | 2nd year College | 62.0000 | 9.3333 |
| Employed | Post grad with units | 61.0000 | 8.0000 |
| Unemployed | 2nd year PS PS/N-T/TV | 60.0000 | 1.0000 |
| Employed | Grade 3 | 58.0000 | 8.1667 |
| Unemployed | College graduate | 55.0000 | 3.2000 |
| Employed | Grade 6 | 54.0000 | 8.2963 |
| Employed | College graduate | 50.0000 | 8.2222 |
| Employed | Grade 8 | 48.6000 | 8.7000 |

As can be seen here, the group with the oldest beneficiaries are the employed group whose highest educational attainment is fifth grade.

*4.1.1. Roll-up*

If this data is too comprhensive, we can roll up on either of the columns. Let both be demonstrated.

Rolling up on job index, we get the following query.

SELECT educal.desc AS educal ,AVG(age) AS avgAge,AVG(work\_ddhrs) AS avgWorkHrs

FROM (SELECT \* FROM ARCDP A INNER JOIN household H ON A.household = H.id) X INNER JOIN educal ON X.educal = educal.id

GROUP BY educal

ORDER BY avgAge DESC,avgWorkHrs DESC

Which yields 24 rows, the top 10 of which are in Table 2.

**Table 2 - ARCDP Beneficiaries Classified by Highest Educational Attainment**

|  |  |  |
| --- | --- | --- |
| **Highest Educational Attainment** | **Average Age** | **Average Work Hours** |
| 4th year College or higher | 63.0000 | 16.0000 |
| 2nd year PS PS/N-T/TV | 60.0000 | 5.7500 |
| Post grad with units | 54.5000 | 2.6667 |
| College graduate | 53.1250 | 5.5789 |
| Grade 6 | 48.1667 | 6.3784 |
| 2nd year College | 47.7500 | 5.1429 |
| Grade school graduate | 47.0000 | 4.0000 |
| Grade 9/(3rd Year HS | 43.8000 | 519.0000 |
| High school graduate | 41.5333 | 5.0526 |
| Grade 10/(4th Year HS | 40.6667 | 2.8750 |

From this data, we can see that the highest average age belongs to the group of people whose highest educational attainment is 4th year college. The group that works the most, however, is the group whose highes educational attainment is Grade 9 or 3rd year high school.

Rolling up on Highest Educatoinal Attainment, we get the following query:

SELECT jobind.desc AS jobind ,AVG(age) AS avgAge,AVG(work\_ddhrs) AS avgWorkHrs

FROM (SELECT \* FROM ARCDP A INNER JOIN household H ON A.household = H.id) X INNER JOIN jobind ON X.jobind = jobind.id

GROUP BY jobind

ORDER BY avgAge DESC,avgWorkHrs DESC

Which results in 2 rows shown in Table 3.

**Table 3 - ARCDP Beneficiaries Classified by Job Index**

|  |  |  |
| --- | --- | --- |
| **Job Index** | **Average Age** | **Average Work Hours** |
| Employed | 49.5610 | 56.7143 |
| Unemployed | 30.2656 | 0.7838 |
| undefined | 2.2500 | 0.0000 |

It can be seen here that employed beneficiaries are much older than the unemployed ones, which may indicate that younger people in their focus group may tend to be unemployed.

*4.1.2. Drill-down*

Any of the other ten dimensions of the ARCDP query can be utilized. This operations shall be demonstrated with the owner tenure and work class dimensions.

Drilling down from the original query on owner tenure, we get the following:

SELECT tenur.desc AS tenur ,jobind.desc AS jobind ,educal.desc AS educal ,AVG(age) AS avgAge,AVG(work\_ddhrs) AS avgWorkHrs

FROM (SELECT \* FROM ARCDP A INNER JOIN household H ON A.household = H.id) X INNER JOIN tenur ON X.tenur = tenur.id INNER JOIN jobind ON X.jobind = jobind.id INNER JOIN educal ON X.educal = educal.id

GROUP BY tenur,jobind,educal

ORDER BY avgAge DESC,avgWorkHrs DESC

Which results in 86 rows, 10 of which have been included in Table 4.

**Table 4 - ARCDP Beneficiaries Classified by Owner Tenure, Job Index and Highest Educational Attainment**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Owner Tenure** | **Job Index** | **Highest Educational Attainment** | **Average Age** | **Average Work Hours** |
| Living in a public space without rent | Employed | Grade 4 | 69.0000 | 8.0000 |
| Owner, owner-like possession of house and lot | Employed | Grade 5 | 69.0000 | 7.3333 |
| Owner, owner-like possession of house and lot | Employed | 4th year College or higher | 63.0000 | 16.0000 |
| Owner, owner-like possession of house and lot | Employed | 2nd year College | 62.0000 | 9.3333 |
| Rent-free house and lot with consent of owner | Unemployed | Grade 4 | 62.0000 | 0.0000 |
| Owner, owner-like possession of house and lot | Employed | Post grad with units | 61.0000 | 8.0000 |
| Owner, owner-like possession of house and lot | Unemployed | 2nd year PS PS/N-T/TV | 60.0000 | 1.0000 |
| Owner, owner-like possession of house and lot | Employed | Grade 3 | 58.0000 | 7.0000 |
| Owner, owner-like possession of house and lot | Employed | 3rd year College | 55.0000 | 9.0000 |
| Living in a public space without rent | Employed | Grade 6 | 55.0000 | 8.0000 |

It can be seen here that their oldest beneficiaries are living in a public space without rent, but employed and have achieved at least grade 4.

Drilling down further on work class, the resultant query is :

SELECT tenur.desc AS tenur ,workcl.desc AS workcl ,jobind.desc AS jobind ,educal.desc AS educal ,AVG(age) AS avgAge,AVG(work\_ddhrs) AS avgWorkHrs

FROM (SELECT \* FROM ARCDP A INNER JOIN household H ON A.household = H.id) X LEFT JOIN tenur ON X.tenur = tenur.id LEFT JOIN workcl ON X.workcl = workcl.id LEFT JOIN jobind ON X.jobind = jobind.id LEFT JOIN educal ON X.educal = educal.id

GROUP BY tenur,workcl,jobind,educal

ORDER BY avgAge DESC,avgWorkHrs DESC

Which returns 98 rows, 10 of which are in Table 5.

**Table 5 - ARCDP Beneficiaries Classified by Owner Tenure, Work Class, Job Index and Highest Educational Attainment**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Owner Tenure** | **Work Class** | **Job Index** | **Highest Educational Attainment** | **Average Age** | **Average Work Hours** |
| Owner, owner-like possession of house and lot | Working for government/corporation | Employed | High school graduate | 75.0000 | 8.0000 |
| Owner, owner-like possession of house and lot | Working for private household | Employed | Grade 5 | 69.0000 | 8.0000 |
| Living in a public space without rent | Self-employed without any employee | Employed | Grade 4 | 69.0000 | 8.0000 |
| Owner, owner-like possession of house and lot | Self-employed without any employee | Unemployed | College graduate | 67.5000 | 7.3333 |
| Owner, owner-like possession of house and lot | Working for private business/establishment/farm | Employed | 4th year College or higher | 63.0000 | 8.0000 |
| Owner, owner-like possession of house and lot | Self-employed without any employee | Employed | 2nd year College | 62.0000 | 8.0000 |
| Rent-free house and lot with consent of owner | undefined | Unemployed | Grade 4 | 62.0000 | 0.0000 |
| Owner, owner-like possession of house and lot | Working for private household | Employed | Grade 3 | 61.0000 | 8.0000 |
| Owner, owner-like possession of house and lot | Working for private household | Employed | Post grad with units | 61.0000 | 8.0000 |

From this data, it could be deduced that their oldest beneficiaries are government employees who own their house and lot, are employed, and are at least high school graduates.

*4.1.3. Dice and Slice*

From the original query, we can dice and slice the data further on the unused dimensions. For the purposes of discussion, let the dimensions house type and job status be used.

Let the original query be diced and sliced where house type is a multiunit residential. The query becomes

SELECT jobind.desc AS jobind ,educal.desc AS educal ,AVG(age) AS avgAge,AVG(work\_ddhrs) AS avgWorkHrs

FROM (SELECT \* FROM ARCDP A INNER JOIN household H ON A.household = H.id) X LEFT JOIN jobind ON X.jobind = jobind.id LEFT JOIN educal ON X.educal = educal.id

WHERE house\_type = '3'

GROUP BY jobind,educal

ORDER BY avgAge DESC,avgWorkHrs DESC

Which results in 1 row as shown in Table 6.

**Table 6 - ARCDP Beneficiaries Classified by Job Index and Highest Educational Attainment Where House Type is Multiunit Residential**

|  |  |  |  |
| --- | --- | --- | --- |
| **Job Index** | **Highest Educational Attainment** | **Average Age** | **Average Work Hours** |
| Employed | 3rd year College | 55.0000 | 10.0000 |

This result shows that there is only one group of beneifciaries where the house type is a multiunit residential and that is the employed group who attained at least 3rd year college and their average age is 55.

Now dicing and slicing the result on job status where job status is working on different jobs day to day, the query becomes

SELECT jobind.desc AS jobind ,educal.desc AS educal ,AVG(age) AS avgAge,AVG(work\_ddhrs) AS avgWorkHrs

FROM (SELECT \* FROM ARCDP A INNER JOIN household H ON A.household = H.id) X LEFT JOIN jobind ON X.jobind = jobind.id LEFT JOIN educal ON X.educal = educal.id

WHERE jstatus = '3'

GROUP BY jobind,educal

ORDER BY avgAge DESC,avgWorkHrs DESC

Which results in 5 rows, as shown in Table 7.

**Table 7 - ARCDP Beneficiaries Classified by Job Index and Highest Educational Attainment Where The Beneficiary Works On Different Jobs**

|  |  |  |  |
| --- | --- | --- | --- |
| **Job Index** | **Highest Educational Attainment** | **Average Age** | **Average Work Hours** |
| Employed | High school graduate | 75.0000 | 8.0000 |
| Employed | 4th year College or higher | 63.0000 | 8.0000 |
| Employed | Grade 8 | 52.0000 | 5.0000 |
| Employed | Grade 4 | 44.0000 | 6.5000 |
| Employed | Grade 3 | 44.0000 | 5.0000 |

**5. RESULTS AND ANALYSIS**

Testing was done on three aspects of the study: the model, the ETL, and the application.

**5.1. Dimensional Model**

The model was appropriate for OLAP. The original model was in 2NF, which meant it required additional joins in order to support OLAP. The new model was downgraded to 1NF for the arcdp fact table. The crop table and land\_parcel tables were only retained as 2NF because they referred to the same household data. The resultant schema still only required 1 join for these two fact tables and no joins at all for the arcdp table. This allowed processing to be faster. In order to prove this, let sample queries be tested in each fact table.

First, for ARCDP, let a sample OLAP query be taken from Section 4.1.2.

SELECT tenur.desc AS tenur ,workcl.desc AS workcl ,jobind.desc AS jobind ,educal.desc AS educal ,AVG(age) AS avgAge,AVG(work\_ddhrs) AS avgWorkHrs

FROM (SELECT \* FROM ARCDP A INNER JOIN household H ON A.household = H.id) X INNER JOIN tenur ON X.tenur = tenur.id INNER JOIN workcl ON X.workcl = workcl.id INNER JOIN jobind ON X.jobind = jobind.id INNER JOIN educal ON X.educal = educal.id

GROUP BY tenur,workcl,jobind,educal

ORDER BY avgAge DESC,avgWorkHrs DESC

Let this query’s performance be compared to the equivalent query in the original schema. The old schema took 9.5687 seconds average out of 10 trials due to having to perform a join. The new schema only took an average of 0.0469, a 99.5099% improvement in performance.

Secondly, for the crop table, let the following sample query be tested.

SELECT wall.desc AS wall ,roof.desc AS roof ,house\_type.desc AS house\_type ,AVG(crop\_vol) AS avgCrop,SUM(crop\_vol) AS totalCrop,COUNT(\*) as cropCount

FROM (SELECT \* FROM crop A INNER JOIN household H ON A.household = H.id) X LEFT JOIN wall ON X.wall = wall.id LEFT JOIN roof ON X.roof = roof.id LEFT JOIN house\_type ON X.house\_type = house\_type.id

GROUP BY wall,roof,house\_type

ORDER BY avgCrop DESC,totalCrop DESC,cropCount DESC

In the old schema, this query took an average of 12.0278 seconds while in the new schema, it took an average of 3.4462, a 71.3478% performance boost.

Finally, for the land parcel table, let the following query be tested.

SELECT tenur.desc AS tenur ,wall.desc AS wall ,roof.desc AS roof ,AVG(alp\_area) AS avgArea,SUM(alp\_area) AS totalArea,COUNT(\*) AS landCount

FROM (SELECT \* FROM land\_parcel A INNER JOIN household H ON A.household = H.id) X LEFT JOIN tenur ON X.tenur = tenur.id LEFT JOIN wall ON X.wall = wall.id LEFT JOIN roof ON X.roof = roof.id

GROUP BY tenur,wall,roof

ORDER BY avgArea DESC,totalArea DESC,landCount DESC;

From these three test cases, it can be deduced that the usage of the data warehousing schema is more efficient than the usage of the original schema, due to the lower normalization standard, which requires lesser joins and processes only the necessary data, which reduces the size of the intermediate tables.

**5.2. ETL Process**

A simple way to verify the validity of the ETL process is to check if a query performed on the old database would yield the same results as a query in the new schema.

Let the same queries be used.

First, let the arcdp table be tested. The old schema results in 98 rows, as well as the new schema. However, the results are quite different. This is due tot he schema’s enforced key integrity. The old CBMS schema allowed duplicate entries in the database. In fact, there are 13 duplicate entires in the household table from executing the query.

SELECT id, COUNT(\*)

FROM db\_hpq.hpq\_hh

GROUP BY id

HAVING COUNT(\*) > 1

As such, the new schema did not allow these entries to be processed at all, which is why the results are marginally different.

Secondly, executing the crop query on both schemas yields 82 rows, again, with marginal differences due to the enforced key integrity.

Finally, executing the land\_parcel query on both schemas yields 178 rows, also with marginal differences.

These marginal differences are small enough, however, that it does not change the sort order on any of the averages of the measures, so meaningful data can still be extracted from the transformed schema.

**5.3. Application**

As mentioned in the introduction, the goal of this project is to realize the use and importance of data warehousing, and the reasons why it is not a trivial matter that isn’t useful. Companies and organizations such as MASIPAG, PRRM, LEARN, and NCCO in particular can make good use of a good OLAP schema rather than use the original, for reasons such as speed (a well-normalized OLAP schema requires lesser joins, allowing analytical queries to execute faster).

The OLAP schema used in this Major Course Output in particular is useful to the aforementioned non-governmental organizations for the following reasons:

*5.3.1. Magsasaka at Siyentipiko para sa Pag-unlad ng Agrikultura (MASIPAG)*

As a network of  NGOs working towards the sustainable use of agricultural resources, MASIPAG would benefit from this OLAP, particularly on the CROP part of the application. A simple query they can do is

SELECT tenur.desc AS tenur ,concat(mun,',',zone,',',brgy,',',purok) AS location\_id ,water.desc AS water ,croptype.desc AS croptype ,AVG(crop\_vol) AS avgCrop,SUM(crop\_vol) AS totalCrop,COUNT(\*) as cropCount

FROM (SELECT \* FROM crop A INNER JOIN household H ON A.household = H.id) X LEFT JOIN tenur ON X.tenur = tenur.id INNER JOIN location L ON X.location\_id = L.id LEFT JOIN water ON X.water = water.id LEFT JOIN croptype ON X.croptype = croptype.id

GROUP BY tenur,location\_id,water,croptype

ORDER BY avgCrop DESC,totalCrop DESC,cropCount DESC

From this query, one can easily get a sense of the types of conditions under which farmers in Marinduque and Palawan live. For example, while not the majority, there are a number of farmers who live on lands without consent of the owner.

*5.3.2. Philippine Rural Reconstruction Movement (PRRM)*

The PRRM is a non-governmental organization formed in 1952, based on rural reconstruction experiences in China in the early 1900s, with the goal of assisting the poorer members of society in the Philippines. Their goal means that the OLAP schema is useful to them because our database contains data about the housing and location of people in Marinduque and Palawan. For example, here is a sample query drilling down on the housing details of people in the database, namely, Wall Type, Roof Type, and House Type:

SELECT wall.desc AS wall ,roof.desc AS roof ,house\_type.desc AS house\_type ,COUNT(\*) AS benefCount,AVG(age) AS avgAge,AVG(work\_ddhrs) AS avgWorkHrs

FROM (SELECT \* FROM ARCDP A INNER JOIN household H ON A.household = H.id) X LEFT JOIN wall ON X.wall = wall.id LEFT JOIN roof ON X.roof = roof.id LEFT JOIN house\_type ON X.house\_type = house\_type.id

GROUP BY wall,roof,house\_type

ORDER BY benefCount DESC,avgAge DESC,avgWorkHrs DESC

This query can be utilized in order to determine who needs help with reconstruction, since good housing can help these people deal with calamities better. From the data, it appears that the majority of people have strong houses, or at least those people who are ARCDP members.

*5.3.3. Labor Education and Research Network (LEARN)*

LEARN is a non-governmental organization that provides various services primarily to workers, whether informal or formal in the nature of their labor. Founded in October 1986, its core programs are education, research, publications, solidarity, and networking.

This one is a little vague or broad in its core programs, however, its services are geared towards workers. First off, knowing the details of work of the people they’re catering to would help (so, querying the columns about work and jobs). Also, one of their core programs is education, so having firsthand knowledge of the highest educational attainment of the people, as well as their location, would help determine who first to help. In drilling down on the Highest Educational Attainment column, it appears that the majority of people in Palawan and Marinduque have not graduated high school.

*5.3.4. National Cooperation of Cooperatives (NCCO)*

This organization is the largest cooperative federation in the Philippines. Among these cooperatives are the Marinduque Social Action Multipurpose Cooperative, and the SRT Narra Cooperative of Palawan. For these people, an all-around knowledge of as many provinces would best suit their goal, however, assuming that the OLAP schema caters to all 77 provinces and municipalities that the NCCO is involved in, they would find relevant data that would help their goals.

**6. CONCLUSION**

For Online Analytical Processing, it is important to have a schema specifically designed for the purpose, an ETL process which, as accurately as possible, transforms the Online Transaction Processing database or mined data into an OLAP schema, and an objective for the OLAP system to have, all of which was successfully accomplished by the team.

By denormalizing an OLTP database schema to 1NF or at best, 2NF, the OLAP schema would require lesser joins, allowing analytical queries, which are heavier than transactional ones, to execute faster.

The ETL process must be accurate in order to reflect the same data as the original OLTP database, only with a more efficient schema. Marginal errors are acceptable provided the dataset is large enough for these errors not to affect any decisions influenced by the system.

The queries generated must be relevant, like the system’s queries were to MAISPAG, PRRM, LEARN, and NCCO in making decisions regarding crop volume, land area, and ARCDP beneficiaries.

Overall, the team was successfully able to implement an OLAP system that has an efficient schema, a correct ETL process, and relevant queries.The team recommends using Microsoft SQL Server or some other DBMS that supports data cubes in order to have faster queries when it comes to OLAP query processing.

**7. REFERENCES**

[1] About MASIPAG | Masipag.org. (n.d.). Retrieved March 16, 2016, from http://masipag.org/about-masipag/

[2] Labor Education and Research Network – LEARN. (n.d.). Retrieved March 16, 2016, from http://www.learn.org.ph/

[3] NATCCO Network - About Us. (n.d.). Retrieved March 16, 2016, from http://www.natcco.coop/index.php/about-us

[4] 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.

[5] Types of Variables. (n.d.) Retrieved March 12, 2016, from Laerd Statistics: https://statistics.laerd.com/statistical-guides/types-of-variable.php

[6] University of Wisconsin Center for Cooperatives - Cooperative Grocer - Philippine Cooperative Movement. (n.d.). Retrieved March 16, 2016, from http://www.uwcc.wisc.edu/info/abroad/sibal.html