National Science Foundation:
Division of Institution & Award Support

INTERNAL TRACKING DATABASE

Syed Sheraz | Harlin Sanders | Lei Wang
ISTM 6202| 30 April 2017

Mission Statement

A portion of the federal government relies on the Division of Institution & Award Support at the National Science Foundation (NSF) to review project proposals and award applications for various research projects. With a wide range of scientific objectives, the goals of these projects could benefit not only our country but also humanity on a global scale. Based on information contained within these proposals, the NSF will conduct scientific feasibility reviews as well as financial reviews to determine award eligibility and size. These reviews are conducted by staff members of various departments within the NSF. While the current system is all paper based, the NSF needs to modernize by digitalizing the documents and information needs and storing electronic data to promote transparency and tracking as well as to save money by increasing efficiencies within these processes.

Requirements

The NSF database should modernize processes by digitizing the documents and information needs to facilitate an electronic trail of these reviews and award applications, increase transparency by providing management with tools to measure results, increase productivity by creating efficiencies, and reduce time spent chasing papers.

The NSF research project funding database will keep track of institutions submitting proposals and seeking funding, Principal Investigators who perform research at the proposing institution, the award they seeking, the department who receives the project proposal and requests financial reviews of the proposal, and the analyst who conducts the financial review.

- For each institution, the NSF will track the TaxID, name, congressional district, street address, zip code, point of contact, and phone number(s) for that point of contact.
- For each Principal Investigator, the NSF will track their first name, last name, educational background, type of science, institution, and investigator ID.
- For each award, the NSF will track the award ID, start date, end date, award amount and award title.
- For each department, the NSF will track the department ID, approving official, branch chief, phone, and science section.
- For each financial review, the NSF will track review ID, type, start date, and will derive end date.
- For each analyst, the NSF will track their first name, last name, analyst ID, and phone number(s).
- A principal investigator is an employee an institution that performs research and proposes the idea for a proposal/grant application.

- The institution applies submits the proposal and award application to the NSF. The award is forwarded to the NSF who conducts a scientific feasibility study and financial review.
- An analyst at the appropriate NSF department is assigned to conduct a financial review.

Conceptual Design: ER Diagram with EER Superclass/Subclass Entities:

Based on the mission statement and requirements, the team collaborated to build the following entity relationship (ER) diagram. The diagram addresses these requirements by using many-to-many relationships (for example, between the ANALYST entity and the FINANCIALREVIEW entity and enhanced entity relationship (EER) superclass/subclass entities (using the INSTITUTION entity as the superclass entity and EDUCATION and CORPORATION as the subclass entities)
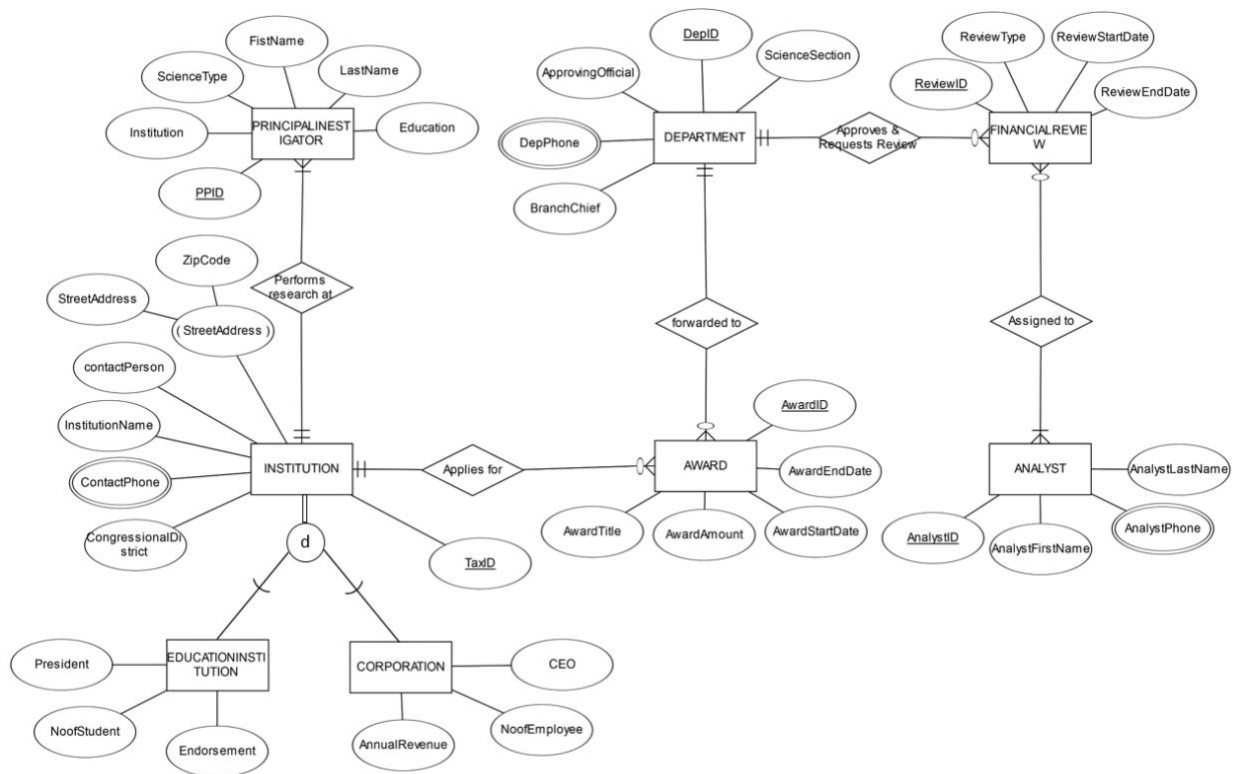
Logical Design: Relational Schema

After instructor feedback and team discussion, the ER diagram was then converted into the below relational schema. This schema was developed using ERDPlus and also includes referential integrity constraint lines.
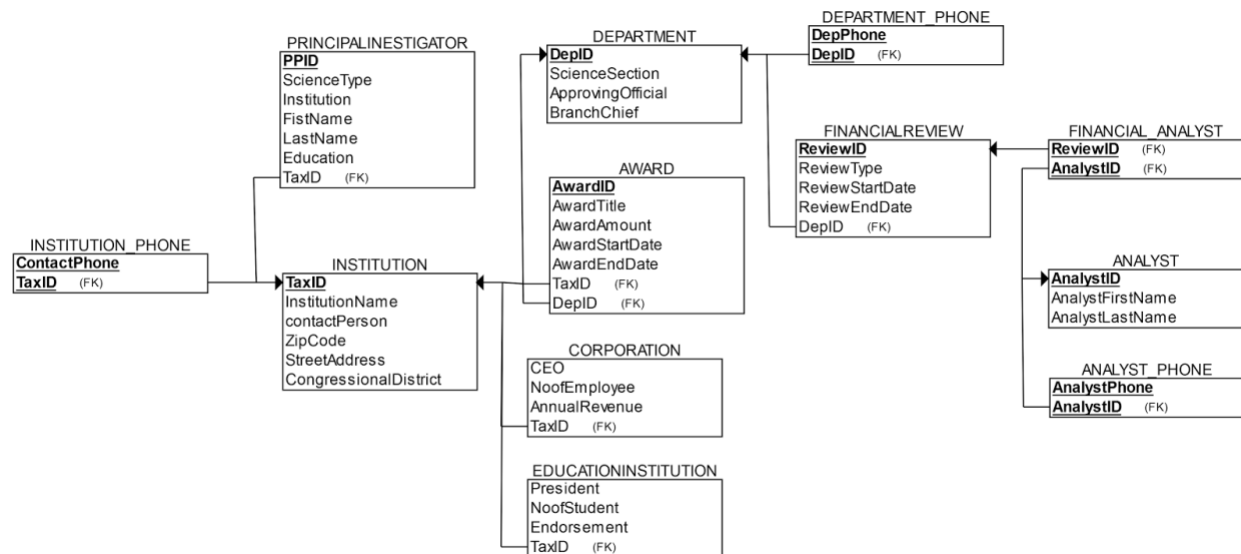
Implementation:

Based on feedback and team discussion, the relational schema was implemented utilizing MySQL on an Amazon RDS instance. The instance, located at istm6202.cdocnyekhdps.us-west- 2.rds.amazonaws.com:3306, has a master username of Group3 and a master password of ISTM6202group3.

The following screenshot depicts the tables created to implement our database.

```
+----------------------+
| Tables_in_project    |
+----------------------+
| ANALYST              |
| ANALYST_PHONE        |
| AWARD                |
| CORPORATION          |
| DEPARTMENT           |
| DEPARTMENT_PHONE     |
| EDUCATIONINSTITUTION |
| FINANCIALREVIEW      |
| FINANCIAL_ANALYST    |
| INSTITUTION          |
| INSTITUTION_PHONE    |
| PRINCIPALINVESTIGATOR |
+----------------------+
```

Queries:

To formulate queries, the team envisioned three primary groups of users of the database and their unique views: 1) analysts; 2) first-line managers; and 3) department branch chiefs.

Analysts

The analysts conduct the financial reviews. Therefore, we created queries (some of which contain complex subqueries) to capture typical questions and analysis performed by the analysts.

Analysts would be interested in grouping the reviews by institution in order to gain efficiencies. By pairing awards from the same institution and then accessing basic organizational and institutional information, analyst could save time and resources to conduct the financial review by eliminating unnecessarily duplicative research.

The following query asks "What institutions have applied for multiple awards?" and it provides the institution's tax ID, name, contact person, street address and zip code as well as congressional district:

SELECT TaxID, InstitutionName, ContactPerson, StreetAddress, ZipCode, CongressionalDistrict FROM INSTITUTION WHERE TaxID IN (SELECT TaxID from AWARD GROUP BY TaxID HAVING COUNT(*)>1);

TaxID	InstitutionName	ContactPerson	StreetAddress	ZipCode	CongressionalDistrict
001	Villanova	Kerry Kittles	800 E Lancaster	19085	7

Again, in order to gain efficiencies and to capitalize on previously-conducted basic research, the following query provides the tax ID, name, contact person, street

address, zip code, and congressional district of each institution with an average award size of over $35,000:

> select TaxID, InstitutionName,ContactPerson, StreetAddress, ZipCode, CongressionalDistrict FROM INSTITUTION WHERE TaxID IN (select TaxID from AWARD group by TaxID HAVING AVG(AwardAmount) > 35000);

```
+-------+-----------------+---------------+----------------------------+---------+----------------------+
| TaxID | InstitutionName | ContactPerson | StreetAddress              | ZipCode | CongressionalDistrict |
+-------+-----------------+---------------+----------------------------+---------+----------------------+
| 001   | Villanova       | Kerry Kittles | 800 E Lancaster            | 19085   |                    7 |
| 005   | AppleTech       | Bob Cousy     | 100 Legends Way            | 20611   |                   10 |
| 012   | Glob            | Garret Hailey | 514 S. Magnolia St. Orlando, FL | 32806   |                   35 |
+-------+-----------------+---------------+----------------------------+---------+----------------------+
```

While not terribly complicated, the most common analyst queries will likely involve accessing an award by title or keyword. Therefore, the following query was created to demonstrate locating an award by the keyword "weapon" in the title:

> SELECT * from AWARD where AwardTitle='Weapon';

```
+---------+-------------------------------+-------------+----------------+--------------+-------+-------+
| AwardID | AwardTitle                    | AwardAmount | AwardStartDate | AwardEndDate | TaxID | DepID |
+---------+-------------------------------+-------------+----------------+--------------+-------+-------+
| 103     | Anthrax as a Bio-Chemical Weapon |       10000 | 2017-03-01     | 2021-03-01   | 003   | 001   |
+---------+-------------------------------+-------------+----------------+--------------+-------+-------+
```

First-line Managers

The first-line managers direct resources (people and money) to conduct financial reviews, evaluates the performance of their unit and the analysts within the unit, and controls the timelines and workflows. conduct the financial reviews. Therefore, we created queries (some of which contain complex subqueries) to capture typical questions and analysis performed by the analysts. Therefore, we created queries (some of which contain complex subqueries) to capture typical questions posed by first-line managers.

To help the first-line manager with resource allocation and human capital management decisions, the following query retrieves the analyst ID, first name, and last name for each analyst currently assigned to more than 3 reviews:

> SELECT ANALYST.AnalystID, AnalystFirstName, AnalystLastName, count(*) as count FROM ANALYST, FINANCIALREVIEW, FINANCIAL_ANALYST WHERE ANALYST.AnalystID = FINANCIAL_ANALYST.AnalystID AND FINANCIALREVIEW.ReviewID = FINANCIAL_ANALYST.ReviewID GROUP BY AnalystFirstName HAVING COUNT > 3 ORDER BY count DESC;

```
+-----------+------------------+-----------------+-------+
| AnalystID | AnalystFirstName | AnalystLastName | count |
+-----------+------------------+-----------------+-------+
| 001       | Chris            | Webber          |     6 |
| 002       | Jalen            | Rose            |     5 |
+-----------+------------------+-----------------+-------+
```

For performance reviews and decisions regarding to whom to assign reviews for processing, the following query provide the analyst ID for any financial analyst who spent more than 30 days on a financial review:

SELECT ANALYST.AnalystID, AnalystFirstName, AnalystLastName, FINANCIALREVIEW.ReviewID, ReviewType, timestampdiff(DAY, ReviewStartDate,ReviewEndDate) as Days
FROM ANALYST, FINANCIALREVIEW, FINANCIAL_ANALYST WHERE ANALYST.AnalystID = FINANCIAL_ANALYST.AnalystID AND FINANCIALREVIEW.ReviewID = FINANCIAL_ANALYST.ReviewID AND timestampdiff(DAY, ReviewStartDate,ReviewEndDate) > 30
GROUP BY AnalystFirstName;

```
+-----------+------------------+-----------------+----------+--------------------+------+
| AnalystID | AnalystFirstName | AnalystLastName | ReviewID | ReviewType         | Days |
+-----------+------------------+-----------------+----------+--------------------+------+
| 010       | Candida          | Wilson          | 006      | Indirect Cost Rate |  245 |
| 001       | Chris            | Webber          | 001      | Audit              |   91 |
| 002       | Jalen            | Rose            | 001      | Audit              |   91 |
| 003       | Jimmy            | King            | 001      | Audit              |   91 |
| 004       | Juwan            | Howard          | 001      | Audit              |   91 |
| 009       | Oswald           | Virgo           | 005      | Indirect Cost Rate |  183 |
| 006       | Piers            | Jameson         | 004      | Audit              |   62 |
| 005       | Ray              | Jackson         | 001      | Audit              |   91 |
| 011       | Reed             | Jephson         | 006      | Indirect Cost Rate |  245 |
| 008       | Sophia           | Simpson         | 005      | Indirect Cost Rate |  183 |
| 007       | Velda            | Walters         | 004      | Audit              |   62 |
+-----------+------------------+-----------------+----------+--------------------+------+
```

In order to allow the first-line manager to see the schedule of reviews being executed and plan for future scheduling and resources, the following query provides the analyst IDs, first, and last names of the analysts conducting financial reviews along with their start and end dates. Additionally, it returns the department ID and science section affiliated with each of these reviews and groups the results first by review ID and then by analyst ID.

SELECT F.ReviewID, A.AnalystID, A.AnalystFirstName, A.AnalystLastName, R.ReviewStartDate, R.ReviewEndDate, R.DepID, D.ScienceSection FROM ANALYST A, FINANCIAL_ANALYST F, DEPARTMENT D, FINANCIALREVIEW R WHERE A.AnalystID = F.AnalystID AND F.ReviewID = R.ReviewID AND R.DepID= D.DepID ORDER BY ReviewID, A.AnalystID;

```
+----------+-----------+----------------+---------------+-----------------+---------------+-------+----------------+
| ReviewID | AnalystID | AnalystFirstName | AnalystLastName | ReviewStartDate | ReviewEndDate | DepID | ScienceSection |
+----------+-----------+----------------+---------------+-----------------+---------------+-------+----------------+
| 001      | 001       | Chris          | Webber        | 2015-09-01      | 2015-12-01    | 002   | Chemistry      |
| 001      | 002       | Jalen          | Rose          | 2015-09-01      | 2015-12-01    | 002   | Chemistry      |
| 001      | 003       | Jimmy          | King          | 2015-09-01      | 2015-12-01    | 002   | Chemistry      |
| 001      | 004       | Juwan          | Howard        | 2015-09-01      | 2015-12-01    | 002   | Chemistry      |
| 001      | 005       | Ray            | Jackson       | 2015-09-01      | 2015-12-01    | 002   | Chemistry      |
| 002      | 001       | Chris          | Webber        | 2012-01-01      | 2012-03-01    | 003   | Geology        |
| 002      | 002       | Jalen          | Rose          | 2012-01-01      | 2012-03-01    | 003   | Geology        |
| 002      | 004       | Juwan          | Howard        | 2012-01-01      | 2012-03-01    | 003   | Geology        |
| 003      | 001       | Chris          | Webber        | 2013-01-15      | 2013-07-15    | 005   | Physics        |
| 003      | 002       | Jalen          | Rose          | 2013-01-15      | 2013-07-15    | 005   | Physics        |
| 003      | 003       | Jimmy          | King          | 2013-01-15      | 2013-07-15    | 005   | Physics        |
| 003      | 005       | Ray            | Jackson       | 2013-01-15      | 2013-07-15    | 005   | Physics        |
| 004      | 001       | Chris          | Webber        | 2014-12-01      | 2015-02-01    | 001   | Biology        |
| 004      | 002       | Jalen          | Rose          | 2014-12-01      | 2015-02-01    | 001   | Biology        |
| 004      | 003       | Jimmy          | King          | 2014-12-01      | 2015-02-01    | 001   | Biology        |
| 004      | 006       | Piers          | Jameson       | 2014-12-01      | 2015-02-01    | 001   | Biology        |
| 004      | 007       | Velda          | Walters       | 2014-12-01      | 2015-02-01    | 001   | Biology        |
| 005      | 001       | Chris          | Webber        | 2015-06-01      | 2015-12-01    | 004   | Infoscience    |
| 005      | 008       | Sophia         | Simpson       | 2015-06-01      | 2015-12-01    | 004   | Infoscience    |
| 005      | 009       | Oswald         | Virgo         | 2015-06-01      | 2015-12-01    | 004   | Infoscience    |
| 006      | 001       | Chris          | Webber        | 2014-03-15      | 2014-11-15    | 006   | MaterialScience |
| 006      | 002       | Jalen          | Rose          | 2014-03-15      | 2014-11-15    | 006   | MaterialScience |
| 006      | 010       | Candida        | Wilson        | 2014-03-15      | 2014-11-15    | 006   | MaterialScience |
| 006      | 011       | Reed           | Jephson       | 2014-03-15      | 2014-11-15    | 006   | MaterialScience |
+----------+-----------+----------------+---------------+-----------------+---------------+-------+----------------+
```

In order to allow the first-line manager to plan for future scheduling and resources, the following query provides the range of financial review timelines (including the shortest and longest duration).

SELECT ReviewID, ReviewType, timestampdiff(DAY, ReviewStartDate,ReviewEndDate) as Days FROM FINANCIALREVIEW GROUP BY ReviewID, ReviewType DESC;

```
+----------+--------------------+------+
| ReviewID | ReviewType         | Days |
+----------+--------------------+------+
| 002      | New Awardee Review |   60 |
| 004      | Audit              |   62 |
| 001      | Audit              |   91 |
| 003      | Over 10 M Review   |  181 |
| 005      | Indirect Cost Rate |  183 |
| 006      | Indirect Cost Rate |  245 |
+----------+--------------------+------+
```

Department Branch Chiefs

The department branch chiefs (for each respective science section) conducts scientific feasibility reviews and approves and requests financial reviews.

Because the department branch chiefs need to request the appropriate resources (people and money) to conduct feasibility reviews, they need to understand the number of awards being submitted for review. Therefore, the following query answers the question, "Which specific department has the most awards assigned to it?":

SELECT DEPARTMENT.DepID, ScienceSection, count(*) as count FROM DEPARTMENT, AWARD WHERE DEPARTMENT.DepID = AWARD.DepID GROUP BY ScienceSection ORDER BY count DESC LIMIT 3;

```
+-------+---------------+-------+
| DepID | ScienceSection | count |
+-------+---------------+-------+
| 001   | Biology       |     5 |
| 003   | Geology       |     2 |
| 005   | Physics       |     2 |
+-------+---------------+-------+
```

To enable the department branch chiefs to ensure appropriate resources are available and request additional resources, the following query considers the departments who have more than one award, provide the department ID, award amount, award title, award start and end dates for each award over $20,000. Group these results by department ID.

SELECT DepID, AwardAmount, AwardTitle, AwardStartDate, AwardEndDate FROM AWARD WHERE AwardAmount >= 20000 GROUP BY DepID HAVING COUNT(*) > 1;

```
+-------+-------------+------------------------------------------------------+----------------+--------------+
| DepID | AwardAmount | AwardTitle                                           | AwardStartDate | AwardEndDate |
+-------+-------------+------------------------------------------------------+----------------+--------------+
| 001   |       50000 | The Cellular Reproduction of DNA                     | 2015-09-01     | 2016-09-01   |
| 002   |       25000 | Geotechnical Properties of Lateritic Soils in Osun State | 2015-06-01 | 2019-06-01   |
| 003   |     1000000 | Big Data: Anonymity, Privacy, and Ethics             | 2016-05-15     | 2021-05-15   |
+-------+-------------+------------------------------------------------------+----------------+--------------+
```

To efficiently report progress to NSF senior management, department branch chiefs will also be able to determine all awards (title, amount and department) that concluded in the previous year (2016 in this case) by utilizing the following single query:

SELECT * FROM AWARD WHERE AwardEndDate LIKE '%2016%';

```
+---------+----------------------------------------------------+-------------+----------------+--------------+-------+-------+
| AwardID | AwardTitle                                         | AwardAmount | AwardStartDate | AwardEndDate | TaxID | DepID |
+---------+----------------------------------------------------+-------------+----------------+--------------+-------+-------+
| 101     | The Cellular Reproduction of DNA                   |       50000 | 2015-09-01     | 2016-09-01   | 001   | 001   |
| 102     | Why Do Rodents Spread Disease and Virus So Quickly? |       25000 | 2014-03-15     | 2016-03-15   | 002   | 001   |
| 111     | Intranet Optimization                              |       10000 | 2015-06-01     | 2016-06-01   | 007   | 004   |
| 114     | Observations of recent solar eclipses              |       25000 | 2015-11-25     | 2016-11-25   | 010   | 008   |
+---------+----------------------------------------------------+-------------+----------------+--------------+-------+-------+
```

Most Sophisticated Queries

Of the queries above, the five most sophisticated queries for our database include the following:

SELECT ANALYST.AnalystID, AnalystFirstName, AnalystLastName, count(*) as count FROM ANALYST, FINANCIALREVIEW, FINANCIAL_ANALYST WHERE ANALYST.AnalystID = FINANCIAL_ANALYST.AnalystID AND FINANCIALREVIEW.ReviewID = FINANCIAL_ANALYST.ReviewID GROUP BY AnalystFirstName HAVING COUNT > 3 ORDER BY count DESC;

SELECT ANALYST.AnalystID, AnalystFirstName, AnalystLastName, FINANCIALREVIEW.ReviewID, ReviewType, timestampdiff(DAY,

ReviewStartDate,ReviewEndDate) as Days
FROM ANALYST, FINANCIALREVIEW, FINANCIAL_ANALYST WHERE
ANALYST.AnalystID = FINANCIAL_ANALYST.AnalystID AND
FINANCIALREVIEW.ReviewID = FINANCIAL_ANALYST.ReviewID AND
timestampdiff(DAY, ReviewStartDate,ReviewEndDate) > 30
GROUP BY AnalystFirstName;

SELECT F.ReviewID, A.AnalystID, A.AnalystFirstName, A.AnalystLastName,
R.ReviewStartDate, R.ReviewEndDate, R.DepID, D.ScienceSection FROM
ANALYST A, FINANCIAL_ANALYST F, DEPARTMENT D, FINANCIALREVIEW
R WHERE A.AnalystID = F.AnalystID AND F.ReviewID = R.ReviewID AND
R.DepID= D.DepID ORDER BY ReviewID, A.AnalystID;

SELECT DEPARTMENT.DepID, ScienceSection, count(*) as count FROM
DEPARTMENT, AWARD WHERE DEPARTMENT.DepID = AWARD.DepID
GROUP BY ScienceSection ORDER BY count DESC LIMIT 3;

SELECT TaxID, InstitutionName, ContactPerson, StreetAddress, ZipCode,
CongressionalDistrict FROM INSTITUTION WHERE TaxID IN (SELECT TaxID
from AWARD GROUP BY TaxID HAVING COUNT(*)>1);

Reports:

These primary user groups were also considered by the team when creating the three
following reports to access anticipated common questions/queries:

Report 1: Award Submissions by Department

Department ID	Award ID	Award Title
001	101	The Cellular Reproduction of DNA
001	102	Why Do Rodents Spread Disease and Virus So Quickly?
001	103	Anthrax as a Bio-Chemical Weapon
001	104	Effects of Drinking Moderate to High Levels of Coffee
001	115	Geology Structure Influence on Petroleum Exploit
Total Number of Awards in Department 1 (Biology): 5		

Department ID	Award ID	Award Title
002	105	Geotechnical Properties of Lateritic Soils in Osun State
002	112	Improvement of Bearing Capacity of Sandy Soil by Grouting
Total Number of Awards in Department 2 (Chemistry): 2		

Department ID	Award ID	Award Title
003	112	Radioactive Elements Research
003	108	Big Data: Anonymity, Privacy, and Ethics
Total Number of Awards in Department 3 (Geology): 2		

Department ID	Award ID	Award Title
004	111	Intranet Optimization
Total Number of Awards in Department 4 (InfoScience): 1		

Department ID	Award ID	Award Title
005	110	Oxide Film Growth for Magnetic Applications
005	113	Newtonian Mechanics and Magnetic Field
Total Number of Awards in Department 5 (Physics: 2		

Department ID	Award ID	Award Title
008	114	Observations of Recent Solar Eclipses
Total Number of Awards in Department 5 (Astronomy): 1		
Total Number of Awards in All Departments: 13		

Repo

Analyst ID
001
001
001
001
Tot...

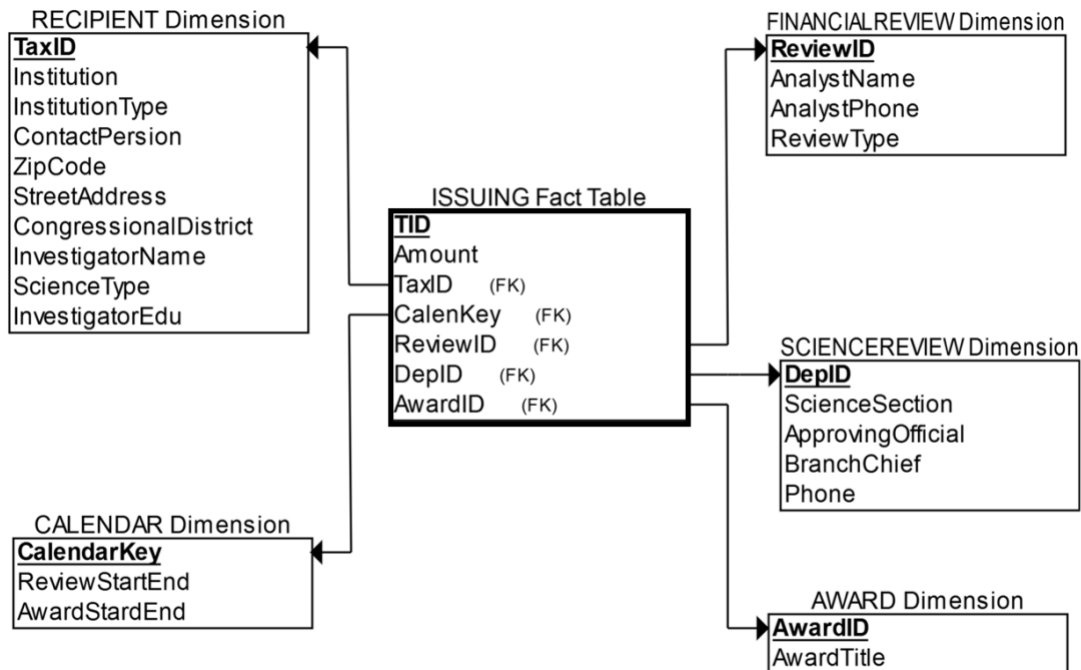| Analyst ID |
|---|
| 002 |

Tax ID

001

001

Tax ID

002

Transforming the Database into a Data Warehouse:

With growth of the data resources in the database, it is potentially necessary for the organization to develop a Data Warehouse and transform existing data into the Data Warehouse. We suggest the organization to take following steps to implement this process:

- Designing Phase: Requirements are collected by the team and visualized for the organization to review and improve the model. An implementable model is then created according to the finalized requirements. The dimensional model, often consist of star schema and ER diagrams, will guide the development throughout subsequent development stages.
- Creating Phase: An ETL infrastructure will be created to extract data from source database, transform relational data into fact tables and populate the data warehouse in deployment phase. Front-end applications are also developed for indirect-end users to read and analyze the data.
- Deployment Phase: Dimension tables will be created, ETL will conduct the transformation work and input into the fact tables with data extracted from source database. ETL infrastructure will continue to populate the data warehouse with updated data or from new data sources. Continuous improvement on performance and functions of the data warehouse will be performed to ensures it matches the BI requirements.

Specifically, a Data Warehouse could be created by utilizing core elements of the initial database and creating a new aggregated fact table, thereby creating a star schema (as seen below). One such aggregated fact table could be developed for use by NSF senior management to better summarize the awards issued by time to each recipient that clearly tracks the review, science section, and title. The aggregation could enable NSF officials to lobby for additional congressional funding as well as ensure that similar projects are appropriately deconflicted across the entire foundation.

RECIPIENT Dimension
TaxID
Institution
InstitutionType
ContactPersion
ZipCode
StreetAddress
CongressionalDistrict
InvestigatorName
ScienceType
InvestigatorEdu

ISSUING Fact Table
TID
Amount
TaxID   (FK)
CalenKey   (FK)
ReviewID   (FK)
DepID   (FK)
AwardID   (FK)

FINANCIAL REVIEW Dimension
ReviewID
AnalystName
AnalystPhone
ReviewType

SCIENCE REVIEW Dimension
DepID
ScienceSection
ApprovingOfficial
BranchChief
Phone

CALENDAR Dimension
CalendarKey
ReviewStartEnd
AwardStardEnd

AWARD Dimension
AwardID
AwardTitle

Project Evolution and Improvement:

Throughout the project, the team continued to make refinements as new techniques were learned and additional topics were discussed in the course. Instructor feedback was also welcomed and influenced a number of key changes.

Specifically, the initial ER diagram created in Phase I of the project was examined and changes were made to address the additional requirements of the project assignment. The changes are as follows. After aggressively pursuing ways to implement enhanced entity relationships by exhaustive research and engaging the system develops of the instructor-recommended ERDPlus program, the team was able to include EER superclass/subclass entities on the INSTITUTION table. This table, implemented as the superclass, was broken down into subclasses of Education and Corporation subclasses given both types of organization can submit NSF proposals. As one can clearly see from the above ER diagram, these subclass entities are distinguished by their own sets of unique attributes. Education institutions, for example, have attributes such as president, number of students, and endorsements/endowments. Whereas Corporation institutions have attributes such as CEO, number of employees, and annual revenue.

A continuous review cycle was created by the team whose members regularly revisited each entity, attribute, and relationship to ensure for an efficient program that addresses the needs of the NSF.