جمهورية العراق

وزارة التعليم العالي والبحث العلمي

جامعة ديالى

تصميم مستودع بيانات ديناميكي باستخدام خوارزمية اليراعات وخوارزمية اسراب الطيور الكمية

رسالة

مقدمة الى قسم علوم الحاسوب / كلية العلوم/ جامعة ديالى وهي جزء من متطلبات نيل درجة الماجستير في علوم الحاسوب

من قبل

وصال اديب عبدالله

بإشراف

أ. ناجي مطر سحيب **أ.م. د. جمال مصطفى عباس**

1440 هـ **2019م**

Chapter One

Introduction

Chapter One

Introduction

1.1 Introduction

Today, large enterprise have huge amount of data and need to analyze these data daily basis. Traditional online transactional processing (OLTP) system becomes inadequate to meet the needs for deep analysis of multi-dimensional data because the high-speed increase in data volume and the complexity of data queries. Here they proposed the concept of relational model and online analytical processing (OLAP), therefore data warehouse become increasingly the key element in the enterprise information technology architecture [1].

Data warehouse (DW) is the storage of large historic data which is collected from multiple operational databases to support complex queries from decision support systems. These complex queries demand summaries data and want minimum response time to produce the result. High response time results when running queries on data warehouse. To reduce the query response time and made faster access of data they use materialized views rather than the source data in data warehouse materialized views are pre-computed views storing in data warehouse [2].

Data warehouse is also an approach for integrating data from multiple heterogeneous databases and other information sources. DW can be seen as a collection of materialized views defined over a group of base relations and when these base relations change the materialized views need to be updated [3].

Data warehouse involves a series of processes that turn source data into data suitable to be analyzed. These series of processes called ETL (extraction, transformation, loading) exports data from various data sources in different formats and after these process load them into the data warehouse [4].

The most important decisions in designing data warehouse is selecting a set of derived views to materialize which minimize the total query response time and maintenance cost of the selected views [5].

Selection of materialized views depends on the user's requirements (e.g. , frequently used queries, query processing and storage cost). Materialized view like a data cache which is a copy of data from distributed data warehouse that can be retrieved quickly [6].

The most important tools in data warehouse are OLAP (on-line analytical processing) and Data mining. OLAP organize data as a multidimensional model suitable for analyzing by the analyzers while Data mining perform the analysis on the data and provide the results to the decision support system. Therefor OLAP support Model-Driven analysis while Data mining support Data-Driven analysis [7].

1.2 Literature Review

Below is a review of some researches which are related to our work:

- **A. N .M. B. Rashid, and M. S. Islam (2010) [8],** This research proposed a methodology to determine whether a view is useful or not for materialization based on many factors like: complexity, view selectivity and database size for object-relational database management system (ORDBMS). They calculate the materialized view maintenance cost using re-materialization and incremental maintenance and conclude that the incremental materialized view maintenance is better than re-materializing the views in ORDBMS.
- **P. P. Karde and Dr. V. M. Thakare (2010) [9],** This research proposed tree based materialized view selection algorithm for query processing and node selection algorithm for fast materialized view selection in distributed environment based on many parameters like: storage space, cost of query, net benefit and cost of maintenance. They evaluate the total cost of different query patterns and frequencies using three different view materialization strategies: all virtual views method, all materialized views method and proposed materialized views method, the results improve that the total cost evaluated from using the proposed materialized views method was smallest among the three strategies. The proposed algorithm minimizes the total time of query processing because it requires shortest total processing time.
- **A. Mohod and M. Chaudhari (2013) [6],** This research selected a set of materialized views according to various parameters like: frequency of query processing, cost of query processing and storage space. The proposed system determines which query is more beneficial for creation the materialized view to obtain high query performance. The proposed framework is executed on

the simulated student data warehouse model using list of query to find the efficiently of the proposed approach in selection of materialized view.

- **S. Choudhary, and R. Mahajan (2014) [10],** they presented methodology to achieve high query performance by determining which queries are more beneficial for the creation of materialized views. To find the efficiency of the designed approach they executed the designed framework on the customer data warehouse model using list of queries. They depend on the access time required to determine the performance of any query through selected materialized views with respect to those query submitted directly to data warehouse. The time required of query through the materialized view is very much less as compared to that query directly selected by data warehouse. The experimental result shows that the performance of queries in terms of access time by selected materialized views was found 15 milliseconds while those queries directly selected from data warehouse was found 32 milliseconds. They also preserve the selected materialized views depending upon the area required to be stored by using mathematical model.
- **D. Yao et. al (2015) [1],** This research improved the cost model and the polynomial greedy algorithm presented in the past for addressing the storage space constraints of the databases to select the views which have overall minimum cost from the candidate views to materialize them. The polynomial greedy algorithm (PGA) works in two stages. In the first stage, it selects the set of views as a candidate views and in the second stage, it selects the views of the largest gains to be materialize. This algorithm calculates the available space firstly and according to that it decides adding or deleting the candidate materialized views by selecting a lower cost for selecting views. The experimental results approved that the algorithm was effective.

- **J. Rajurkar, and T. K. Khan (2015) [11],** This paper developed Bitmap pruning strategy for speed up query processing since index pruning based approach eliminate the need of scanning and processing the whole dataset. The results show that the bitmap index saving disk access by avoiding tuple scan on a table with more number of attributes and also it reduces computation time by processing bitwise operations. The researchers also developed vector alignment algorithm to overcome the problem of massive empty AND results by using priority queues.
- **C. D. Garhwani et. al (2016) [12],** This research presented a comprehensive review on processing large datasets. They presented two techniques: aggregate function based technique and compressed bitmap based technique with World Aligned Hybrid (WAH) compression technique. They developed various compression schemes for compressing the bitmap index and proposed new compression technique called variable length compression technique to overcome the pitfalls of existing systems. They observed that compressed bitmap index using variable length coding is efficient for data warehouse query processing and OLAP since bitmap index has the following benefits: 1) saving disk access by avoiding tuple scan on a table with more number of attributes. 2) Reducing computation time by conducting bitwise operations.

1.3 Problem statement

Data Warehouse has a large number of tables that are connected with each other in relationship. Each table has a large number of columns and rows up into millions. Data Warehouse grows continuously since new information is being generated continuously by operational systems. New complex queries can submit on data warehouse, for this reason big database need an efficient way that be able to answer database queries, make query response time better and increase system performance.

1.4 Aim of thesis

The aims of this thesis are:

1- Selecting optimal materialized views which have high access frequency besides low processing time and storage space.
2- Decreasing the complex queries response time and optimizing its performance by posing these queries to the summary tables (optimal frequent materialized views) instead of the base tables.

1.5 Thesis Organization

This thesis contains four chapters addition to chapter one.

- **Chapter Two: Theoretical Background**

  This chapter clarifies the definition, architecture and the design of data warehouse in addition to the multidimensional model, and some applications of data warehouse like Decision Support Systems (DSSs) and Online

Analytical Processing (OLAP), it also explain the summarized and bitmap index technique in data warehouse and give a brief description about algorithms used to select best materialized views. Finally it gives the basics of the relational algebra.

- **Chapter Three: The Proposed System**

  This chapter illustrates the main steps of the proposed system for designing dynamic data warehouse.

- **Chapter Four: Implementation results and evaluation**

  This chapter shows the implementation results of the proposed system step's and evaluate these results.

- **Chapter Five: Conclusions and Suggestions for Future work**

  This chapter presents the conclusions of the proposed work and many suggestions for the future works.

الخلاصة

نمت كمية المعلومات في المؤسسات الكبيرة في السنوات الاخيرة بشكل انفجاري, تولد النظم التشغيلية معلومات جديدة بشكل مستمر, جلب هذا تحدي جديد في عملية صنع القرار المبني على هذه الكميات الهائلة من البيانات, مستودع البيانات يمكن ان ينظر إليه على انه مجموعة من الجداول و وجهات النظر المعرفة على هذه الجداول بمجموعة من العلاقات, وجهات النظر هذه يمكن ان تعرف على انها دوال مشتقة لجداول ناتجة من مجموعة من الجداول الاساسية ويعاد حسابها في كل مرة تطلب فيها. القضية الحساسة في تصميم مستودع البيانات هي استخدام بعض الميكانيكيات مثل الجداول الملخصة او ماتعرف بوجهات النظر المتحققة و الفهارس لغرض زيادة سرعة معالجة الاستعلام, بسبب قيود المساحة الخزنية من المستحيل تحقيق جميع وجهات النظر في مستودع البيانات لذلك اقترح العديد من الباحثين طرق مختلفة تهدف لحل هذه المشكلة بواسطة اختيار افضل مجموعة من وجهات النظر و تحقيقها في مستودع البيانات.

اقترحت هذه الرسالة منهجية كفوءة لتصميم مستودع بيانات ديناميكي بواسطة اختيار افضل وجهات النظر المتحققة بصورة دورية, لتحسين اداء الاستعلامات المعقدة و تقليل وقت استجابتها, يتضمن العمل المقترح اربع مراحل رئيسية التي هي: تحميل مستودع البيانات المطلوب و تطبيق دوال التجميع بحالتين (عبر الوصول المباشر للجداول الاساسية و عبر مؤشر النقطية) و حساب كلف الاختيار لوجهات النظر المتحققة و ذلك بحساب المعلمات (وقت المعالجة, المساحة الخزنية وتردد الوصول) لكل وجهة نظر متحققة ثم تطبيق التقنية الهجينة (خوارزمية اليراع و خوارزمية اسراب الطيور الكمية) لأختيار افضل وجهات النظر المتحققة التي تمتلك اعلى تردد وصول و مساحة خزنية اقل بالاضافة الى وقت معالجة اقل.

بينت نتائج العمل المقترح ان مؤشر النقطية حقق نتائج جيدة لاستعلامات التجميع حيث ان وقت استجابة هذه الاستعلامات على احد الحقول في مستودع البيانات المستخدم كان 757 ملي ثانية بينما وقت استجابة هذه الاستعلامات نفسها على نفس الحقل عبر مؤشر النقطية كان 489 ملي ثانية, و بهذا فأن اداء استعلامات التجميع عبر مؤشر النقطية افضل من الوصول المباشر عبر الجداول الاساسية بأكثر من 54%, قللت التقنية الهجينة الوقت الكلي لتكوين افضل وجهات النظر المتحققة لأن خوارزمية اليراع قدمت نقطة ابتدائية مثالية لخوارزمية اسراب الطيور الكمية.