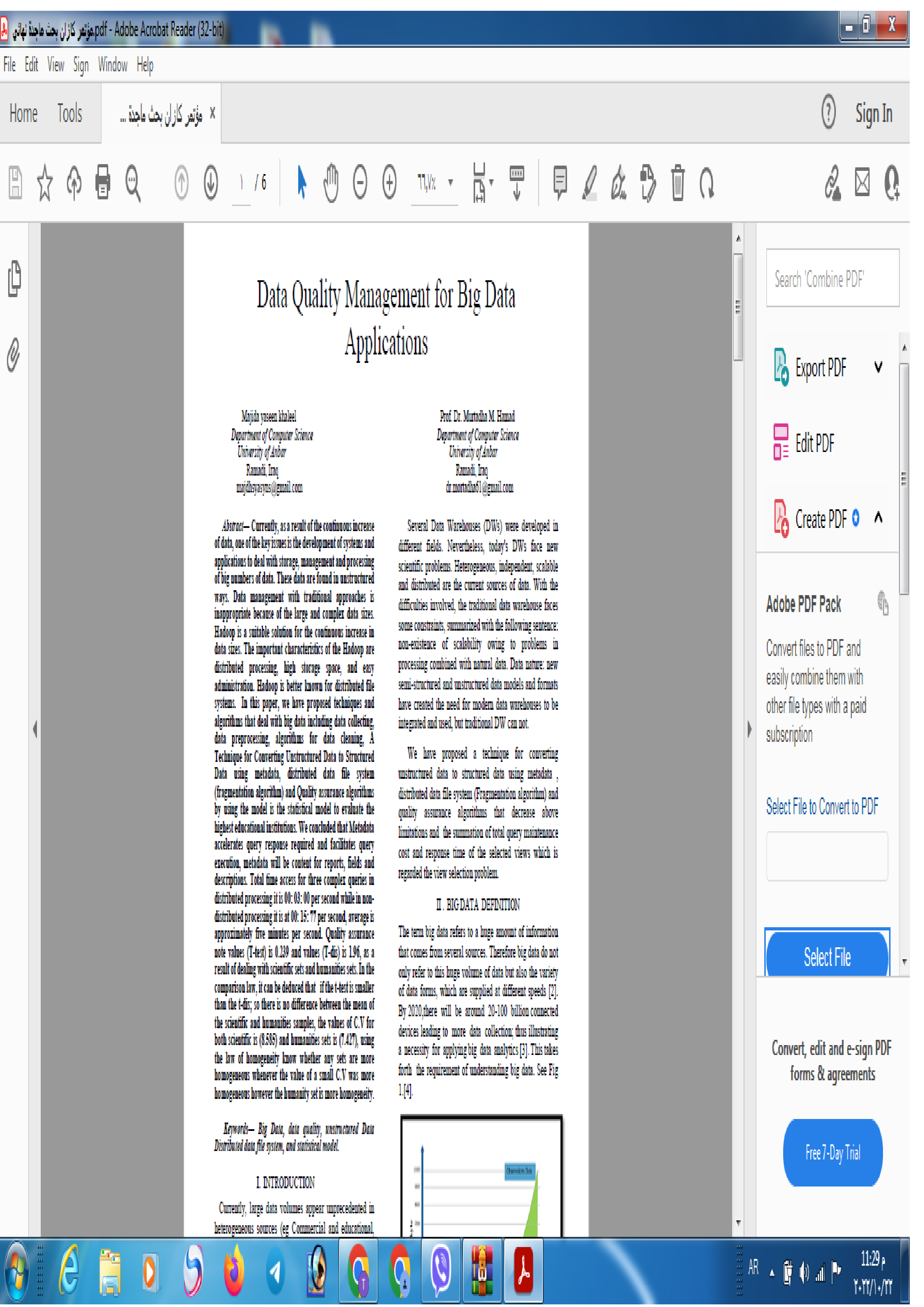
مؤتمر كازان بحث ماجدة نهائي.pdf - Adobe Acrobat Reader (32-bit)
File Edit View Sign Window Help
Home Tools
مؤتمر كازان بحث ماجدة ...
Sign In
1 / 6
Search 'Combine PDF'
Export PDF
Edit PDF
Create PDF
Select File
Free 7-Day Trial
AR
11:29 م
٢٠٢٢/١٠/٢٢

Data Quality Management for Big Data Applications

Majida yaseen khaleel
Department of Computer Science
University of Anbar
Ramadi, Iraq
majidhsyasyns@gmail.com

Prof. Dr. Murtadha M. Hamad
Department of Computer Science
University of Anbar
Ramadi, Iraq
dr.mortadha61@gmail.com

Abstract— Currently, as a result of the continuous increase of data, one of the key issues is the development of systems and applications to deal with storage, management and processing of big numbers of data. These data are found in unstructured ways. Data management with traditional approaches is inappropriate because of the large and complex data sizes. Hadoop is a suitable solution for the continuous increase in data sizes. The important characteristics of the Hadoop are distributed processing, high storage space, and easy administration. Hadoop is better known for distributed file systems. In this paper, we have proposed techniques and algorithms that deal with big data including data collecting, data preprocessing, algorithms for data cleaning, A Technique for Converting Unstructured Data to Structured Data using metadata, distributed data file system (fragmentation algorithm) and Quality assurance algorithms by using the model is the statistical model to evaluate the highest educational institutions. We concluded that Metadata accelerates query response required and facilitates query execution, metadata will be content for reports, fields and descriptions. Total time access for three complex queries in distributed processing it is 00: 03: 00 per second while in non-distributed processing it is at 00: 15: 77 per second, average is approximately five minutes per second. Quality assurance note values (T-test) is 0.239 and values (T-dis) is 1.96, as a result of dealing with scientific sets and humanities sets. In the comparison law, it can be deduced that if the t-test is smaller than the t-dis; so there is no difference between the mean of the scientific and humanities samples, the values of C.V for both scientific is (8.585) and humanities sets is (7.427), using the law of homogeneity know whether any sets are more homogeneous whenever the value of a small C.V was more homogeneous however the humanity set is more homogeneity.

Keywords— Big Data, data quality, unstructured Data Distributed data file system, and statistical model.

I. INTRODUCTION

Currently, large data volumes appear unprecedented in heterogeneous sources (eg Commercial and educational,

Several Data Warehouses (DWs) were developed in different fields. Nevertheless, today's DWs face new scientific problems. Heterogeneous, independent, scalable and distributed are the current sources of data. With the difficulties involved, the traditional data warehouse faces some constraints, summarized with the following sentence: non-existence of scalability owing to problems in processing combined with natural data. Data nature: new semi-structured and unstructured data models and formats have created the need for modern data warehouses to be integrated and used, but traditional DW can not.

We have proposed a technique for converting unstructured data to structured data using metadata , distributed data file system (Fragmentation algorithm) and quality assurance algorithms that decrease above limitations and the summation of total query maintenance cost and response time of the selected views which is regarded the view selection problem.

II. BIG DATA DEFINITION

The term big data refers to a huge amount of information that comes from several sources. Therefore big data do not only refer to this huge volume of data but also the variety of data forms, which are supplied at different speeds [2]. By 2020,there will be around 20-100 billion connected devices leading to more data collection; thus illustrating a necessity for applying big data analytics [3]. This takes forth the requirement of understanding big data. See Fig 1.[4].

Adobe PDF Pack

Convert files to PDF and easily combine them with other file types with a paid subscription

Select File to Convert to PDF

Convert, edit and e-sign PDF forms & agreements