KDU Journal of Multidisciplinary Studies (KJMS)
Volume 1. Issue1, 2017

KJMS 2017 VOL.1 (1): 16-24

BIG DATA AND PREDICTIVE ANALYTICS: TIME SERIES FORECASTING FOR IMPROVED DECISION MAKING IN BUSINESS APPLICATIONS

KG Hewa[1]
Department of Information Technology, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka, krishnihw@gmail.com[1]

WPJ Premarathne[2]
Department of Computer Science, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka, punsi11@gmail.com[2]

ABSTRACT

Predictive analytics is a significant aspect of big data analysis where large data sets are analyzed in order to give future predictions. Reliability and accuracy of the information generated through the use of predictive analysis models are crucial, particularly in the business environment. Many business organizations have progressed to integrate predictive analytics to their systems in order to get a competitive advantage in the business arena. Forecasting is critical to the successful execution of strategic as well as operational functions of an organization which further emphasizes the significance of the precision level of the forecasted information. Time series modeling and forecasting is a widely-used technique in monitoring and analyzing industrial processes to generate forecasted business data. This paper critically evaluates and analyses the moving average, exponential smoothing and linear regression predictive analytics models, which are widely used in time series forecasting, determining their reliability and accuracy with the use of business data. The accuracy of the time series forecasting models was further established through forecast error measurement statistics where it was determined that the linear regression model is a better fit for business data. The secular trend and the comparison of the actual data set with the forecasted data sets also helped determine the extent of accuracy of the predictive models.

KEY WORDS: *Big data, predictive analytics models, time series, business data, forecast error measurement statistics*

Corresponding Author: KG HEWA email: krishnihw@gmail.com

1 INTRODUCTION

Predictive models are crucial in the business environment where their results can be employed in order to make critical business decisions that will provide organizations, particularly profit-oriented organizations, with a competitive advantage. Accurate forecasting is vital in making various management decisions for both strategic and tactical planning. Time Series is a forecasting technique where forecasts are made on a collection of data obtained sequentially over a particular period of time. Recent developments in technology along with Internet of Things applications have provided massive volumes of data, which can be analyzed in order to create information that is imperative to the success of business organizations. Big data has become a significant aspect in almost every organization and is vital in the process of analyzing data through Time Series Predictive Models in order to assist the organization in making better management decisions and performing strategic business moves, such as Finance and Risk Management, Operations Management etc., which will impact the growth of the organization in a positive manner. Advancements in data analysis and software capabilities have led to the development of sophisticated systems that offer effective forecasting to anticipate future demands, schedule productions and reduce inventories. Application of predictive analytics is crucial in the technological sector specifically in the development of strategic level business applications such as Executive Support Systems (ESS) that support the decision-making process of the strategic management. Thus, the accuracy and the reliability of the predictive models are vital in establishing the implementation aspect of the software development process, as the forecasted results provided by the application will directly affect the business decisions made by the strategic level management.

This paper critically evaluates and analyses the Time Series Predictive Models; Moving Average Method, Exponential Smoothing and Linear Regression, through the application of big data pertaining to the annual revenues of an organization. Univariate time series data of annual revenues of Walmart (an American Retailing Corporation), over a period of ten years (from year 2000 to 2009) were used as test data in order to get an in-depth understanding of the precision of the predictive models. The extent of accuracy and the reliability of the time series models were established by the comparison of the actual data and the forecasted values, by which the models which generate more accurate forecasts can be established.

2 RELATED WORK

Organizations have realized the importance of using big data in a business environment due to their confrontation with vast data volumes (Banica and Hagiu, 2015). Big data can be defined as high-volume, high- velocity and varying information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making (Kandalkar and Wadhe, 2014). Recent developments in Internet of Things (IoT) applications have further increased the scale of data to an unprecedented level (Tyagi et al., 2015). Careful synchronization and analyzation of big data, which has become the backbone of corporate performance and economic growth, has the ability to improve organizational performance while enabling better risk management (Chase, 2013). It is commonly sought after for building predictive models, where data forecasting is of major statistical significance (Hand, 2009).

A Time series dataset is a set of observations made in the chronological order, where the consequent values are within a predictable range from one another (Bhaskaran,2012). Prediction through the use of Time series data aims at explicit modeling of variable dependencies to forecast the next few values of the series (Esling and Agon, 2012). The ZETA model for bankruptcy classification developed by E.Altman(Altman et al, 1977), one of the earliest predictive analytics business models, and a Decision Support System for sales prediction developed by Kuo and Xue(1998) using fuzzy neural networks, provide the basis for the more advanced and accurate predictive models developed later. Time series data occur naturally in countless domains including medical analysis (Keogh et al., 2001) financial analysis (Zhu and Shasha, 2002) sensor network monitoring (Papadimitriou and Yu, 2006), gene expression analysis (Lin et al., 2008) etc. Time series models, which are grouped under Quantitative forecasting models, analyse the past observations of a time series to generate forecasts. Economic and Sales forecasting, Stock Market Analysis and Quality Control are some of the applications of time

series forecasting models (Gosasang V. et al., 2011) where general pattern or tendencies are taken into consideration. Research has been carried out in economics (Kang 1996, Timmermann and Granger 2004, Shin and Park 2009), in utility forecasting (Conejo et al. 2005, De Gooijer and Hyndman 2006) and in many other areas through the use of time series modeling. Forecasting Decision Support Systems (FDSS), developed specifically for business organizations, integrate managerial judgment, quantitative methods and databases to aid the forecaster in accessing, organizing and analyzing forecasting related data and judgments (Fildes et al., 2006). The reliability and the accuracy of the forecasting methodologies used in such systems are crucial as they directly influence the strategic and operational functions of an organization.

3 METHODOLOGY

Time series predictive models; Moving Average, Exponential Smoothing and Linear Regression, were used in order to obtain forecasted Walmart revenue values for a time period of five years. Forecasted results from the models were further ascertained through the use of forecasted error measurement statistics.

3.1 Moving average method

Moving average method averages the most recent values, where each observation receives the same weight, in a time series to obtain the forecast (y^*_{t+1}) for the next time period.

$$y^*_{t+1} = \frac{y_{t-n+1} + y_{t-n+2} + \cdots + y_t}{n}$$

(1)

Where,
t = time period for which data is collected
y_t = values (business data) in the time series
n = number of observations

The number of Time series values(n) included when executing this function may vary according to the relevance of the time series values, where a small value (preferably between 3-5) is used in situation where only the most recent time series values are considered relevant.

3.2 Exponential smoothing model

Exponential Smoothing is a Time series technique where forecasts can be calculated explicitly, using the weighted average of all the previous actual values of the time series. This technique employs a smoothing constant (W), specifically between 0 and 1, when applying to a set of business data. A set of smoothed values (E_i) are obtained initially, where the actual yearly values from past records are denoted by Y_i.

$$E_1 = Y_1$$
$$E_2 = WY_2 + (1 - W)E_1$$
$$\vdots$$
$$E_n = WY_n + (1 - W)E_{n-1}$$

(2)

The following equation (3) is used to obtain forecasted values, F_{n+1}.

$$F_{n+1} = WY_n + (1 - W)E_n$$

(3)

The Exponential smoothing method eliminates the drawback in the Moving average method where an equal weight is applied on all the data when computing the average.

3.3 Linear regression model

Linear Regression attempts to model the relationship between an explanatory variable and a dependant variable by fitting a linear equation to observed data. Values obtained from the following formulas are required in order to compute the forecasted values for subsequent years,

$$\bar{t} = \frac{\sum t_i}{n}$$

(4)

$$\bar{y} = \frac{\sum y_i}{n}$$

(5)

$$b = \frac{\sum_{i=1}^{n} y_i t_i - n\bar{y}\bar{t}}{\sum_{i=1}^{n} t_i^{\,2} - n\bar{t}^{\,2}}$$

(6)

$$a = \bar{y} - b\bar{t}$$

(7)

Where,
t_i = time period for which data is collected
y_i = values (business data) in the time series
n = number of observations

The forecasted value(y^*) is calculated using,

$$y^* = a + bt$$

(8)

Where a represents the intercept and b represents the slope of the linear trend line respectively.

3.4 Forecast error measurement statistics

Several error measurement statistics were taken into consideration when ascertaining the level of reliability and accuracy of the predictive models. The forecast error which is required when calculating the error measurement statistics is given by,

Forecast error (e_i) =Actual value (a_i) – Forecast
(10)

3.4.1 Mean absolute error(mae)

Mean Absolute Error(MAE) is calculated by averaging the absolute values of forecast errors.

$$MAE = \frac{\sum_{i=1}^{n} e_i}{n}$$

(11)

3.4.2 Mean squared error(mse)

Mean Squared Error is obtained by averaging the squared value of forecast errors.

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n}$$

(12)

3.4.3 Mean absolute percentage error (mae)

Mean Absolute Percentage Error (MAPE) is calculated by averaging the absolute value of percentage forecast errors.

$$MAPE = \frac{\sum_{i=1}^{n} \left(\frac{e_i}{a_i} \times 100\right)}{n}$$

(13)

3.5 Application development and implementation

An application was developed through the Microsoft Visual Studio 2012 platform and Dev Express using the C# language in order to implement the three predictive models.

System Diagnostics was used in order to obtain the execution time of the respective predictive model to obtain a single forecast.

4 RESULTS

The above-mentioned time series predictive models were applied to the same set of data and the corresponding results for each model are shown below.

The forecasting error statistics (MAE, MSE and MAPE), which are critical in establishing the forecast accuracy for each model as well as execution time were also calculated.

4.1 Moving average

A three-point simple moving average method was employed, where n=3, in order to generate forecasted revenue data for Walmart from year 2010 to 2014. A comparative study of the actual and forecasted values is shown below (Table 1).

Table 1 Walmart Revenues: Forecasted values vs. Actual values (Year 2010 to 2014) using Moving Average method

Year	Actual Values ($ million)	Forecasted Values ($ million)
2010	421,395	393,347
2011	446,509	399,855
2012	468,651	399,445
2013	476,294	397,549
2014	485,651	398,950

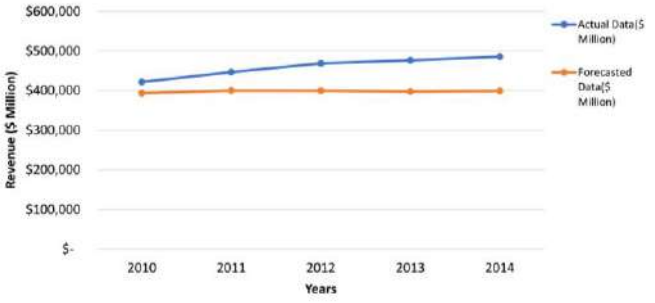

Figure 1 Walmart Revenues: Forecasted values vs. Actual values (Year 2010 to 2014)

According to the graph (Figure 1) there is a significant difference between the forecasted revenues and the actual revenues of Walmart. A slight deviation in the pattern of the forecasted revenue values can be seen for the years 2012 and 2013 whereas the actual revenue values have increased in a linear manner.

Table 2 Forecast Error Statistics for the Moving Average method

MAE	61,871
MSE	4,294,118,976
MAPE (%)	13.25141092

The error measurement statistics for the Moving average method are illustrated above (Table 2).

4.2 Exponential smoothing

A comparison of forecasted values and the actual Walmart revenue values for a time period of five years, from 2010 to 2014 are depicted below (Table 1). A smoothing constant (W) of 0.7 was used for the computation of the forecasting process.

Table 3 Walmart Revenues: Forecasted values vs. Actual values (Year 2010 to 2014) using Exponential Smoothing method

Year	Actual Values ($ million)	Forecasted Values ($ million)
2010	421,395	403,681
2011	446,509	403,377
2012	468,651	403,313
2013	476,294	403,299
2014	485,651	403,297

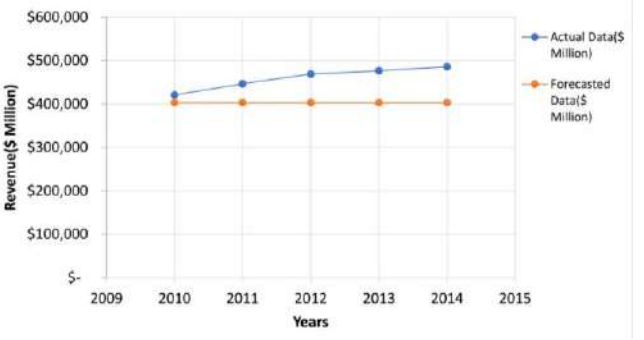

Figure 2 Walmart Revenues: Forecasted values vs. Actual values (Year 2010 to 2014)

The forecasted values seem to have the pattern of decreasing gradually whereas the actual revenues of Walmart have progressed steadily.

Table 4 Forecast Error Statistics for the Exponential Smoothing method

MAE	56,306.6
MSE	3,710,732,161
MAPE (%)	12.01765307

4.3 Linear regression

Forecasted values obtained using the Linear Regression method are shown below.

Table 5 Walmart Revenues: Forecasted values vs. Actual values (Year 2010 to 2014) using Linear Regression method

Year	Actual Values ($ million)	Forecasted Values ($ million)
2010	421,395	445,678
2011	446,509	472,420

Year	Actual Values ($ million)	Forecasted Values ($ million)
2012	468,651	499,162
2013	476,294	525,903
2014	485,651	552,645

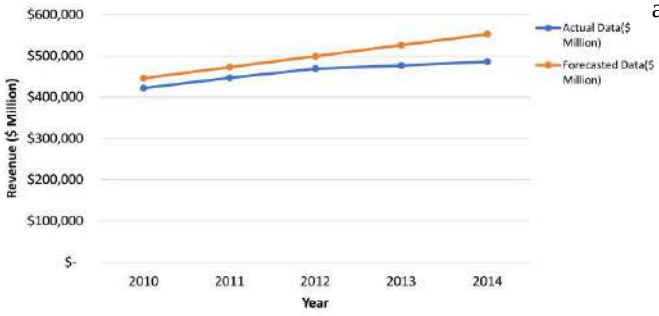

Figure 3 Walmart Revenues: Forecasted values vs. Actual values (Year 2010 to 2014)

The graph (Figure 3) indicates that the forecasted values follow a linear pattern, where the first three forecasted revenue values are very close to the actual revenue values.

Table 6 Forecast Error Statistics for the Exponential Smoothing method

MAE	39,461.6
MSE	1,828,242,810
MAPE (%)	7.72025459

4.4 Execution time analysis

The time taken for each predictive model to generate a single forecast value is mentioned below (Table 7).

Table 7 Time Elapsed for the Computation of one forecast value

Forecast Model	Time Elapsed (Milliseconds)
Moving Average	0.0003
Exponential Smoothing	0.0011
Linear Regression	0.0007

It has to be noted that the computational time for the execution of a single forecast is based on the complexity of the developed algorithm and the number of past observations considered (in this case 10 values from year 2000 to 2009), where CPU utilization speed and memory consumption also comes to play.

5 DISCUSSION

A considerable difference between the sets of forecasted values obtained from the three predictive models and the actual revenue values was observed.

Table 8 Walmart Revenues: Forecasted values vs. Actual values (Year 2010 to 2014)

Year	Actual Values ($ million)	Forecasted Values ($ million)		
		Moving Average	Exponential Smoothing	Regression
2010	421,395	393,347	403,681	445,678
2011	446,509	399,855	403,377	472,420
2012	468,651	399,455	403,313	499,162
2013	476,294	397,549	403,299	525,903
2014	485,651	398,950	403,297	552,645

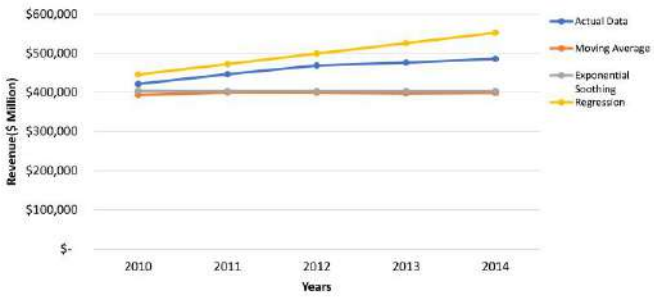

Figure 4 Walmart Revenues: Forecasted values vs. Actual values (Year 2010 to 2014)

As illustrated in figure 4, the forecasted values obtained using the Moving average method and the Exponential smoothing method show similarities in value range and pattern. Although the values forecasted using the Linear regression model are significantly higher than that those of the other models, the forecasted values are slightly higher and somewhat closer to the actual values the secular trend line of the Actual data and the data from the

Linear Regression model show a prominent similarity. Thus, it can be said that comparatively, the Linear regression model provides more accurate results.

Furthermore, the distribution of the set of actual values and the sets of forecasted values derived from the three predictive models, illustrated in the box plot shown below (Figure 5), was also taken into consideration when determining the model that is the better fit for business data.

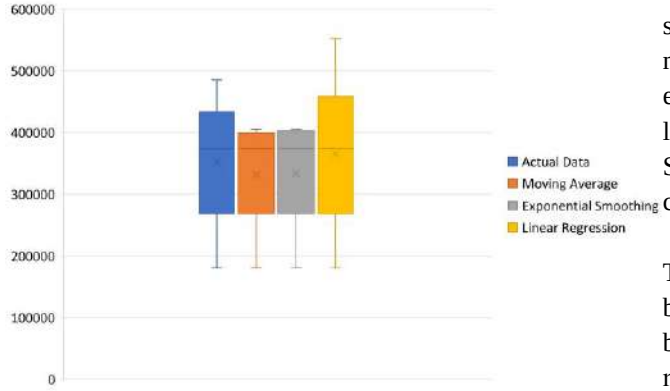

Figure 5 Distribution of Walmart Revenues: Forecasted values and Actual values (Year 2000 to 2014)

The median of all the data sets remain the same while the interquartile ranges of the Actual dataset and the Linear Regression Data set are closer than those of the other data. When considering the distribution of data, the spread of actual data and the data set with the Linear Regression results show similarities. This further elaborates the accuracy and the reliability of the Linear Regression method over the other two forecasting models discussed above.

The forecast error measurement statistics obtained for each predictive model is depicted below (Table 9).

Table 9 Forecast Error Statistics for the Moving Average, Exponential Smoothing and Linear Regression methods

Predictive Model	MAE	MSE	MAPE (%)
Moving Average	61,871	4,294,118,976	13.25141092
Exponential smoothing	56,306.6	3,710,732,161	12.01765307
Linear Regression	39,461.6	1,828,242,810	7.72025459

When considering the forecasted error measurement statistics (MAE, MSE and MAPE) for the predictive models, the statistics for the Moving Average method and the Exponential Smoothing method demonstrate very large values while the statistics for the Linear Regression model are comparatively low. This indicates that the Linear Regression model is a better fit for business data applications than the Moving average and Exponential smoothing methods.

When considering the execution time taken to compute a single forecast (Table 7), the Linear Regression model shows a moderate speed when compared to the other two models. Although the Moving Average shows more efficiency and consumption of less execution time, it has low reliability and accuracy levels, and the Exponential Smoothing model demonstrates low efficiency in comparison to the other models discussed.

The selection of the predictive model, which provides the best fit for business data, is vital in the development of a business application. High reliability and accuracy of the model is extremely crucial as the organization's business strategy is directly influenced by the forecasts provided by the application. The employment of an algorithm based on a predictive model which has a low level of reliability and accuracy, in the development of the business application, will in turn lead to inaccurate forecasts which will be used by the organization to make strategic business decisions yielding catastrophic results. Thus, the establishment of the level of reliability and accuracy of a predictive model is vital when developing business applications that directly influence and support the decision-making process of an organization.

6 CONCLUSION AND FURTHER WORKS

Massive volumes of data have been created and made available to organizations as a result of various business and Internet of Things applications. Time series modelling and forecasting that are crucial in various practical domains employ Big Data to generate various predictions. The extent of accuracy, reliability and efficiency of these predictive models are vital, particularly in the business sector in order to aid in the decision-making process. The selection of a suitable forecasting method that is highly accurate and reliable, to

be implemented in the development of strategic level business applications is significant as the forecasted values provided through the application will directly influence the decision-making process of the strategic level management. The secular trend, value differences and forecasted error measurement statistics for the time series predictive models; Moving average, Exponential Smoothing and Linear regression, were taken in to consideration when determining their extent of accuracy and reliability. Accordingly, the Linear regression model demonstrated a higher level of accuracy comparative to the Moving average model and the Exponential smoothing model, proving to be a better fit for business data. A low execution time was also demonstrated by the Linear Regression model which establishes its high efficiency level.

In the methodology followed, the results obtained were based on secular trend forecasting. The models can be further improved considering Cyclic and Seasonal Variations. The factor of White noise which helps in establishing correlations between variables will further improve the predictive models by determining the relevance of previous values when forecasting future values. Residual diagnostics can also be performed to further determine the reliability and the accuracy of time series forecasting methods. Performance measures, Root Mean Square Error(RMSE) and Theil's U-Statistics can be used to further evaluate and establish forecast accuracy.

7 REFERENCE

Altman, E I; Haldeman, R G; Narayan, P (1977): ZETA Analysis: A new model to identify bankruptcy risk corporations, *Journal of Banking and Finance* 1 (1), 29-54

Banika, L; Hagiu, A (2015): Big Data in Business Environment, *Scientific Bulletin- Economic Sciences* 14 (1), 79-86

Bhaskaran, S (2012): Time series analysis for long term forecasting and scheduling of organizational resources- few cases, *International Journal of Computer Applications* 14 (12), 4-9

Chase, C W (2013): Using Big Data to Enhance Demand-Driven Forecasting and Planning, *Journal of Business Forecasting* 32 (2), 27-32

Conenjo, A J; Plazas, M A; Espinola, R; Nolina, A B (2005): Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA models, *IEEE Transactions on Power Systems* 20 (2), 1035-1042

De Gooijer, J G; Hyndman, R J (2006): 25 Years of Time Series Forecasting, *International Journal of Forecasting* 22 (3), 443-473

Esling, P; Argon C (2012): Time Series Data Mining, *ACM Journal of Computing Surveys* 45 (1),12-45

Fildes, R; Goodwin, P; Lawrence, M (2006): The Design Features of Forecasting Support Systems and Their Effectiveness, *Decision Support Systems* 42 (1), 351-361

Goosang, V; Chanfraprakaikul, W; Kiattisin, S (2011): A Comparison of Traditional and Neural Networks Forecasting Techniques for Container Throughput at Bangkok Port, *The Asian Journal of Shipping and Logistics* 27 (3), 463-482

Hand, D J (2009): Mining the Past to Determine the Future: Problems and Possibilities, *International Journal of Forecasting* 25 (3), 441-451

ICDM (2001): Keogh, EJ; Chu, S; Hart, D; Pazzani, MJ; An online algorithm for segmenting time series. *Proceedings of the 2001 IEEE International Conference on Data Mining*, San Jose, United States of America, 289

Kandalkar, N A; Wadhe, A (2014): Extracting Large Data using Big Data Mining, *International Journal of Engineering Trends and Technology* 9 (11), 576-582

Kang, H (1986): Univariate ARIMA Forecast of Defined Variables, *Journal of Business and Economic Statistics* 4 (1), 81-86

Kuo, R J; Xue, K C (1998): A decision support system for sales forecasting through fuzzy neural networks with asymmetric fuzzy weights, *Decision Support Systems* 24 (2), 105-126

Lin, T; Kaminski, N; Bar-Joseph, Z (2008): Alignment and classification of time series gene expression in clinical studies, *Bioinformatics* 24 (13), 147-155

Miglani, J (2016): *Amazon vs Walmart - Revenues and Profits 1995 to 2015.* [Online] Available at: https://revenuesandprofits.com/amazon-vs-walmart-revenues-profits-1995-2015/, [Accessed: 27th December 2016].

Shin, J; Park, Y (2009): Brownian Agent Based Technology Forecasting, *Technological Forecasting & Social Change* 76 (8), 1078-1091

Sigmond (2006): Papadimitriou, S; Yu, PS; Optimal multi-scale patterns in time series streams, *Proceedings of the ACM SIGMOND International Conference on Management of Data*, Chicago, United States, 647

Timmerman, A; Granger, C W J (2004): Efficient Market Hypothesis and Forecasting, *International Journal of Forecasting* 20 (1), 15-27

Tyagi, A K; Priya, R; Rajeswari, A (2015): Mining Big Data to Predicting Future, *International Journal of Engineering Research and Applications* 5 (3), 14-21

VLDB (2002): Zhu, Y; Shasha, D; StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time, *Proceedings of the 28th International conference on Very Large Databases.* Hong Kong, China, 358