**Ain Shams University**
**Faculty of Engineering**
**Computer and Systems Engineering Department**

# Incremental Data Mining of Data Streams

A Thesis Submitted in Partial Fulfillment of the Requirements of the Degree of Doctor of Philosophy in Computer and Systems Engineering

**Submitted By**

**Amany Fathy Soliman**
M.Sc. In Computer and Systems Engineering
Computer and Systems Engineering Department
Faculty of Engineering, Ain Shams University

**Under Supervision of**

**Prof. Dr. Hoda Korashy Mohammed**
Professor at the Computer and Systems Engineering Department
Faculty of Engineering, Ain Shams University

**Dr. Gamal A. Ebrahim**
Assistant Professor at the Computer and Systems Engineering Department
Faculty of Engineering, Ain Shams University

**Cairo - 2012**

# Acknowledgement

I am heartily thankful to Prof. Dr. Hoda Korashy for her encouragement, guidance, direction and support from the preliminary to the concluding level. I am particularly grateful to Dr. Gamal A. Ebrahim for his thoughtful and creative comments. I am sure it would have not been possible without his help.

I am also grateful to my parents, brother, and sisters who always believe in me.

I also thank my wonderful children: Mahmoud, Mariam, and Menna, for always making me smile. I hope that one day they can read this thesis and understand why I spent so much time in front of my computer.

Finally, words alone cannot express the thanks I owe to Ahmed, my husband, for his unconditional love, encouragement, and support.

Amany Fathy

**Ain Shams University**
**Faculty of Engineering**

# Approval sheet

**Name : Amany Fathy Soliman**

**Degree: Doctor of Philosophy in Computer and Systems Engineering**

**Thesis Title: Incremental Data Mining of Data Streams**

## Discussion Committee

**Prof. Dr. Vijay Raghavan**

Professor at University of Louisiana, United States of America

**Prof. Dr. Hazem Mahmoud Abbas**

Professor at the Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University

**Prof. Dr. Hoda Korashy Mohammed**

Professor at the Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University

Date: 9 / 6 / 2012

# Statement

This dissertation is submitted to Ain Shams University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer and Systems Engineering.

The work included in this thesis was carried out by the author at Computer and Systems Engineering Department, Ain Shams University.

No part of this thesis has been submitted for a degree or a qualification at any other universities or institutions.

Date: 9 / 6 / 2012

Name: Amany Fathy Soliman

Signature:

# Abstract

**Name: Amany Fathy Soliman**

**Degree: Doctor of Philosophy in Computer and Systems Engineering**

**Thesis Title: Incremental Data Mining of Data Streams**

The advances in processing and communication techniques resulted in a multitude of emerging applications that interact with streams of data. Traditional data mining systems store arriving data, collect them for later mining, and make multiple passes over the collected data. Unfortunately, these systems are prohibitively slow when they deal with data streams with massive amounts of data arriving at high rates. Data streams have attracted considerable attention in recent years. A growing number of applications generate streams of data. The continuous generation of new elements in a data stream imposes additional constraints on the methods utilized for mining such data. For example, memory usage is restricted, the infinitely flowing original dataset cannot be scanned multiple times, and current results should be available on demand. In many cases, evolution of sequential patterns is more interesting than sequential patterns themselves. Data evolution is one of the most challenging problems in mining sequential patterns in data streams.

Hence, in this thesis a new framework for mining sequential patterns in evolving data streams is introduced. Batch-window combined with tilted-time window models have been adopted in mining sequential patterns in evolving data streams. Simulation study has been carried out to show the applicability and flexibility of the presented model. The proposed framework guarantees no false negatives and imposes a lower bound of the support of false positives. In addition, the correctness of the proposed framework has been proven.

The introduced framework has been extended to account for distributed data stream situations. The extended model focuses on evolving data streams that originate from multiple distributed sources. Moreover, the mining process is achieved without compromising the privacy of the individual data streams of the participant nodes. The extended framework is able to mine sequential patterns from multiple distributed evolving data streams. It is proven that the proposed model produces no false negatives and imposes a lower bound of the support of false positives. Simulation study has been carried out to analyze the performance of the proposed model. Simulation results show that the proposed model reduces the communication overhead in the distributed mining process compared to performing the mining in a centralized setting. Most importantly, it scales linearly with the number of distributed nodes, which contributes to the scalability of the proposed model.

# Contents

# List of Figures

# List of Tables

# CHAPTER 1

# INTRODUCTION

## 1.1 General

In recent years, data streams have attracted considerable attention in different fields of computer science such as database systems, data mining, and distributed systems. A data stream is an ordered sequence of data items, where the elements of the sequence continuously arrive as time progresses.

A growing number of applications generate streams of data; these applications may include sensor network data, performance measurements in network monitoring and traffic management, and log records generated by web servers.

Because of the underlying resource constraints in terms of memory and running time; most conventional data mining techniques have to be adapted to fit with the nature of the data streams.

## 1.2 Introduction

This chapter provides an overview of the concept of data streams and explains the types and characteristics of data streams. It also, introduces several applications of data stream processing and the challenges that face them. In addition, it presents Data Stream Management Systems (DSMS), continuous queries, and a comparison between DSMS and traditional Database Management Systems (DBMS). The models of data streams will be discussed; in addition, techniques utilized for mining data streams are introduced. Finally, the objectives of the thesis and the scope and organization are presented.

## 1.3 Definition of Data Streams

A data stream is a real-time, continuous, ordered sequence of items [1]. The order of these items is either implicit by arrival time or explicit by timestamps. The order in which items arrive in a data stream could not be controlled and it is not feasible to store a stream locally in its entirety [1]. The items in a data stream arrive at a high rate, which leads to a massive/infinite volume of data. In addition, each item in a data stream is a structured record. In [2] data stream was defined as: “the continuous flow of data generated at a source (or multiple sources) and transmitted to various destinations.”

A variety of areas gives motivation for studying data streams which may include: hundreds of nodes in sensor networks, each taking readings at a high rate; huge quantities of meta-data generated from communications networks about the traffic passing across them; and in