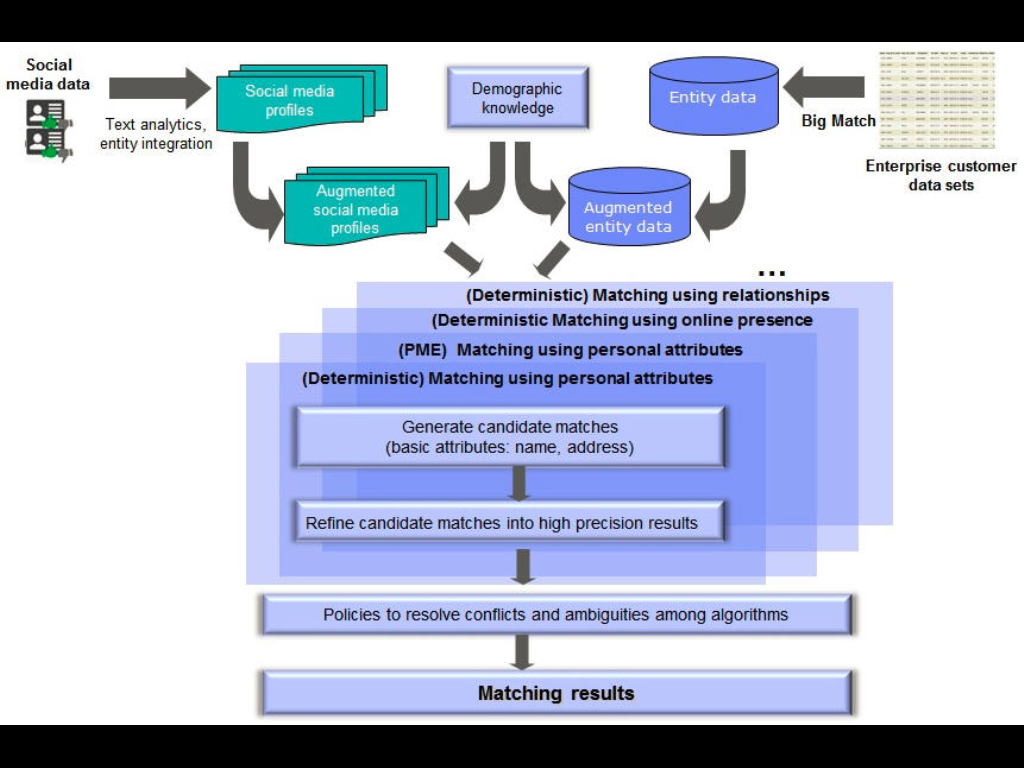
**QDA: QueryDriven Approach to Entity Resolution**

**Abstract:**

This paper explores "on-the-fly" data cleaning in the context of a user query. A novel Query-Driven Approach (QDA) is developed that performs a minimal number of cleaning steps that are only necessary to answer a given selection query correctly. The comprehensive empirical evaluation of the proposed approach demonstrates its signi cant advantage in terms of effeciency over traditional techniques for query-driven applications.The signicance of data quality research is motivated by the observation that the of data-driven technologies such as decision support tools, data exploration,analysis, and scientic discovery tools is closely tied to the quality of data to which such techniques are applied. It is well recognized that the outcome of the analysis is only as good as the data on which the analysis is performed. That is why today organizations spend a substantial percentage of their budgets on cleaning tasks such as removing duplicates,correcting errors, and lling missing values, to improve data quality prior to pushing data through the analysis pipeline.Given the critical importance of the problem, many efforts, in both industry and academia, have explored systematic approaches to addressing the cleaning challenges.

**architecture diagram:**



**EXISTING SYSTEM:**

* These rules can be discovered from existing high quality data such as master data or manually identified data. Inspired by the swoosh method, each cluster is then merged into a composite record via a merge function. Finally a traditional ER method, denoted by T-ER, can be applied to identify the new data set. Moreover, in order to identify more records, the current ER result can be used as the training data to discover new ER-rules. The training data can also be obtained by using techniques, such as relevant feedback, crowd sourcing and knowledge extraction from the web. Therefore, with the accumulated information, ER-rules for more entities can be discovered.
* **Invalid rule.** A rule r is invalid if there exist records that match LHS(r) but do not refer to RHS(r) . Invalid rules might be discovered when the information of entities is not comprehensive. For example, suppose the training data set involves the records. The rule r: (name ¼“wei wang”)^(coa 2“zhang”)) e1 can be generated. For o31, it matches LHS(r) but does not refer to e1. Therefore, r is an invalid rule.
* **Incomplete rule set.** An ER-rule set R of entity set E is incomplete if there are records referring to entities in E that are not covered by R. Both the incomprehensive information of entities and continuous changes of entity features would cause a rule set become incomplete. To solve these problems, we develop some methods to identify candidate invalid rules and candidate useless rules and discover new effective ER-rules.

**PROPOSED SYSTEM:**

Entity resolution is a well-known problem and it has received signicant attention in the literature over the past few decades. A thorough overview of the existing work in this area can be found in surveys . We classify the ER techniques into two categories as follow:Generic ER. A typical ER cycle consists of several phases of data transformations that include: normalization, blocking,similarity computation, clustering, and merging ,which can be intermixed. In the normalization phase, the ER framework standardizes the data formats. The next phase is blocking which is a main traditional mechanism used for improving ER effciency The primary motivation of this paper is querie on online data.A key concept driving the QDA approach is that of vestigiality. A cleaning step (i.e., call to the resolve function fora pair of records) is called vestigial (redundant) if QDA can guarantee that it can still compute a correct answer without knowing the outcome of this resolve. We formalize the concept of vestigiality in the context of a large class of SQL selection queries and develop techniques to identify vestigial cleaning steps.

**System configuration:**

HARDWARE REQUIREMENTS:

Hardware - Pentium IV

Speed - 3.5 GHz

RAM - 1GB

Hard Disk - 20 GB

Key Board - Standard Windows Keyboard

Mouse - Two or Three Button Mouse

Monitor - SVGA

SOFTWARE REQUIREMENTS:

Operating System : Windows

Technology : Java and J2EE

Web Technologies : Html, JavaScript, CSS

IDE : My Eclipse

Web Server : Tomcat

Database : My SQL

Java Version : J2SDK1.8