مقدمه مترجمان

در طـول دهـه هـای گذشـته حجـم زیـادی از داده هـا در پایگـاه داده هـا انباشـته و ذخیـره شـده انـد. امروزه داده ها قلب تپنده فرآیند تجـاری بیشـتر شـرکتها تلقـی میشـوند، آنهـا فـارغ از خـرد وکـلان بـودن نـوع صـنعت در تمـامی صـنایع نظیـر خـرده فروشـها، ارتباطـات، تولیـد، تسـهیلات،حمل و نقـل، بیمـه، کارتهـای اعتبـاری و بانکـداری از طریـق تعـاملات در سیسـتمهای عملیـاتی شـکل مـی گیرنـد. در واقـع سـازمان هـا در اطلاعـات غـرق شـده انـد در حالیکـه تشـنه دانـش هسـتند. داده کـاوی فرآینـد کشـف دانش پنهان درون داده هـا اسـت کـه بـا توصـیف، تشـریح، پـیش بینـی و کنتـرل پدیـده هـای گونـاگون پیرامـونی، دارای کاربردبسیاروسـیعی در حـوزه هـای مختلـف اسـت بـه گونـه ای کـه مـرز و محـدودیتی بـرای کـاربرد آن در نظـر گرفتـه نشـده اسـت. امـروزه، اسـتفاده از روشـهای سـنتی جمـع آوری و تحلیـل داده بـه دلیـل اتـلاف زمـان و ایجـاد هزینـه هـای بسـیار زیـاد، مناسـب نبـوده و ازایـن رو اسـتفاده از روشهای جدید آنالیز داده مانندداده کاوی بسیار حیاتی به نظر میرسد.

معضل اصلی داده کـاوی در ایـران ایـن اسـت کـه بسـیاری از کارشناسـان ایـن حـوزه، بـدون مطالعـه و شـناخت کـافی از ماهیـت داده هـا وقبـل از انتخـاب و پیـاده سـازی بسـتر و متـدولوژی مناسـب جهـت انجام یک پروزه داده کـاوی، بـه سـراغ ابزارهـای داده کـاوی مـی رونـد. بـه نظـر مـی رسـد دلیـل اصـلی این امر عدم رعایت اسـتاندارد کریسـپ درخصـوص نحـوه اجـرا و پیـاده سـازی صـحیح پـروژه هـای داده کاو ی اسـت. ایـن اسـتاندارد در تمـام دنیـا بعنـوان ابـزار مشـترک در پیـاده سـازی و اجـرای پـروژه هـای داده کاوی مورد استفاده کارشناسـان مربوطـه قـرار مـی گیـرد. تهیـه اسـتاندارد فـوق توسـط سـه شـرکت دایمـر کرایسـلر [1] ، SPSS و NCR در سـال ۱۹۹۶ مـورد توافـق قـرار گرفـت . پـیش نـویس اسـتاندارد کریسـپ، اولـین بـار در سـال ۱۹۹۹ نگـارش گردیـد. سـپس کنسرسـیوم CRISP در سـال ۲۰۰۰ نسخه نهـایی را منتشـر نمـود. بـا توجـه بـه اسـتقبال کارشناسـان داده کـاوی از نـرم افـزار SPSS Celementine و اهمیـت اسـتاندارد مـورد اشـاره در پـروژ هـای داده کـاوی، بـرآن شـدیم بـا ترجمـه آخـرین نسـخه از اسـتاندارد فـوق، گـامی مثبـت در راه پیشـرفت و موفقیـت بنگاههـای اقتصـادی و خدماتی از طریق اجرای پروزه های داده کاوی برداریم.

به امید ایران سرشار از ارزش و تهی از اتلاف

[1] Daimler Chrysler

پیشگفتار

CRISP-DM در اواخر سال ۱۹۹۶ توسط سه شرکت با سابقه در زمینه داده کاوی (که در آن ایام پدیده ای جدید و نوظهور بود) تهیه و به بازار عرضه شد. یکی از این شرکت ها Daimler Chrysler بود که بعدها به Daimler-Benz تغییر نام یافت. این شرکت در مقایسه با اغلب بنگاههای تجاری و مراکز صنعتی در زمینه کاربرد نرم افزارهای داده کاوی در عملیات تجاری، تجربه بیشتری داشت.

SPSS (که بعدها به ISL تغییر یافت) از سال ۱۹۹۰ خدمات مبتنی بر عملیات داده کاوی را به متقاضیان عرضه و در سال ۱۹۹۴ اولین نمونه از نرم افزارهای داده کاوی تجاری به نام Clementine را به بازار معرفی نمود.

سومین شرکت با نام NCR ، یکی از اهداف خود را افزایش حجم دیتا یا اطلاعات مورد پردازش برای نرم افزار انبارداری شرکت های طرف قرارداد خود عنوان نمود. پروژه ای که در زمان خود Teradata نام گرفت. این شرکت تیم هایی را به منظور مشاوره در زمینه داده کاوی و افراد متخصص در زمینه فن آوری اطلاعات برای رفع نیاز مشتریان خود، تربیت نمود.

در آن زمان، هر چند استفاده از داده کاوی در عملیات تجاری ناشناخته بود ولی آمارها استفاده گسترده و فراگیر این نوع ابزارها را نشان می داند. این آمارها یا به بیانی دیگر علائم ، برای بعضی خوشنود کننده و برای عده ای نیز ترسناک جلوه می نمود. سعی همه ما بر این بود که در همگامی و بکارگیری ابزارهای داده کاوی، به نحوی، روش ها و تجربیات خود را در اولویت قرار دهیم. دلواپسی ما این بود که، آیا در پیاده سازی این ابزار اشتباه نمی کردیم؟ آیا هر کاربر جدید داده کاوی ، در وهله اول و برای آموختن آن مجبور بود همانند ما از روش سعی و خطا استفاده کند؟ و نیز از منظر یک تامین و تولیدکننده، چگونه می توانستیم مشتریان خود را متقاعد کنیم که داده کاوی ما، از نوع تکامل یافته است و می توان از آن به عنوان یک ابزار کلیدی در تعاملات تجاری استفاده نمود؟

در پاسخ به سوالات بالا و با دلایلی متقاعد شدیم که، ایجاد یک مدل فرایند استاندارد بدون داشتن مالکیت خصوصی (مانند نرم افزارهای متن باز) و با دسترسی بدون هزینه برای کاربر، پاسخگوی ما و تمام کارورزان این نوع نرم افزارها خواهد بود.

یک سال بعد، کنسرسیومی تشکیل دادیم و اسم آنرا با کنار هم گذاشتن حرف اول کلمات :

(Cross-Industry Standard Process for Data Mining)- فرایندهای استاندارد صنعتی برای داده کاوی[1] -ابداع نمودیم. برای تامین بودجه، از کمیسیون اروپایی کمک گرفته و شروع به بازپروری و پیاده سازی ایده های اولیه خود کردیم. از آنجا که هدف از طراحی CRISP-DM ارائه نقش دوگانه آن (یعنی هم از لحاظ ابزاری و هم از لحاظ کاربردی) در صنعت بود، به این نتیجه رسیدیم که تا حد ممکن، از کارورزان یا کاربران بیشتری که با این داده کاوی در ارتباط هستند (مانند مشاوران مدیران ارشد و سرپرستان

[1] Data mining

مدیریت اطلاعات انبارها) جهت جمع آوری اطلاعات مورد لزوم استفاده نماییم. افراد انتخاب شده توسط ما نسبت به فرآیند داده کاوی اطلاعات و دلایل شخصی خود را داشتند. ما برای دستیابی به اهداف خود ، گروهی علاقه مند به کار با این نرم افزار را به اسم (THE SIG) فراهم آوردیم. برای تشکیل این گروه، دعوتنامه هایی را به افراد علاقه مند به حضور در کارگاه آموزشی یک روزه که در شهر آمستردام هلند برگزار می شد ارسال نمودیم.در این کارگاه آموزشی ما توانستیم ایده های خود را برای آنها بازگو نموده و از آنها دعوت نماییم تا نظرات خود را به ما ارائه دهند و راههای تعامل با نرم افزار CRISP-DM را به صورت بحث آزاد مطرح نمایند. در روز برگزاری این کارگاه آموزشی، احساس بیم و هراس در بین اعضای کنسرسیوم به چشم می خورد. یکی از دلواپسی های افراد گروه این بود که، آیا هرگز علاقه مندانی به کار ما وجود خواهند داشت تا در این همایش حاضر شوند؟ و اگر چنین افرادی وجود داشته باشند، آیا به ما خواهند گفت که این نرم افزار به عنوان یک فرآیند یا ابزار استاندارد نیاز قانع کننده ای را جوابگو می باشد یا نه؟ یا اینکه خواهند گفت که ، ایده های ما حتی از ایده ها و تفکرات افرادی که استانداردسازی برایشان یک تخیل غیرعملی است ، بسیار فاصله دارد.

این کارگاه آموزشی از آنچه انتظار می رفت بهتر بود. در آنجا سه چیز جلب توجه می نمود:

- حضور مردم دو برابر تعدادی بود که ما در ابتدای کار انتظار داشتیم.
- وجود یک وفاق عمومی در مورد این موضوع که، "صنایع هم اکنون و بیشتر از هر زمان دیگر نیازمند یک فرآیند استاندارد هستند"، به چشم می خورد.
- همچنان که هر شرکت کننده نظرات و تجربیات شخصی خود را در ارتباط با کاربرد داده کاوی عنوان می نمود(واضح بود که در این اظهارات تفاوت های ظاهری و سطحی وجود دارد مخصوصا در مشخص نمودن حدود فازهای اجرایی و نیز در نوع واژگان فنی بکار برده شده) ولی با این حال، نقطه نظرات مشترک زیادی در دیدگاههای این اشخاص نسبت به فرآیند داده کاوی اطلاعات وجود داشت.

تا زمان به پایان رسیدن این کارگاه آموزشی، این احساس اطمینان در ما بوجود آمد که ما قادر خواهیم بود با استفاده از اطلاعات دریافت شده از گروه (THE SIG) و نیز انتقادات آنها یک مدل استاندارد داده کاوی را به کاربران این عرصه ارائه نماییم.

پس از گذشت دو سال و نیم از برگزاری این کارگاه آموزشی، تلاش ما بر توسعه و پالایش CRISP-DM معطوف شد. ما برای آزمایش این نرم افزار بطور زنده و بصورت یک پروژه بزرگ داده کاوی اطلاعات ، آنرا در شرکت مرسدس بنز و نیز در شرکت بیمه OHRA (از شرکای کنسرسیوم) به اجرا گذاشتیم. ما تلاش نمودیم تا اجزای CRISP-DM را برای استفاده در فضای تجاری هماهنگ و آماده کنیم. گروه SIG ، مفید بودن خود را با رشد در تعداد اعضاء (داشتن بالغ بر ۲۰۰ عضو) و اجرای کارگاههای آموزشی در شهرهای لندن، نیویورک و بروکسل ، بر ما ثابت نمود. تا پایان زمانیکه اتحادیه اروپا نیز به پروژه ما ملحق شد ، یعنی اواسط سال ۱۹۹۹، ما نسخه ای آزمایشی را تولید نمودیم که از لحاظ مدل فرآیند دارای کیفیت بسیار بالایی بود.کسانیکه با این نسخه آشنایی دارند ، تصدیق خواهند نمود که پس از گذشت یکسال ، CRISP-DM – نسخه ۱ با نسخه اولیه خود از لحاظ ظاهر و محتوا تفاوت فاحشی داشت. در واقع ما به

این نکته واقف بودیم که، در طی انجام پروژه ، کار زیادی را برای ارتقاء نسخه اولیه باید انجام دهیم. کارآیی CRISP-DM فقط در گروه کوچکی از پروژه ها (شرکت ها) آزمایش و تایید شده بود. در طی سال گذشته ، شرکت Daimler Chrysler این فرصت را بدست آورد تا CRISP-DM را برای تعداد بیشتری از اپلیکیشن های خود بکار بگیرد.

گروه های متخصص در ارائه خدمات حرفه ای به واحدهای تولیدی و تجاری وابسته به شرکت های SPSS و NCR ، با بکارگیری CRISP-DM توانستند بطور موفقیت آمیزی مشکلات عدیده شرکت های تولیدی و تجاری طرف قرار داد خود را در زمینه خدمات به مشتریان ، حل و فصل نمایند.

در تمام این مدت، ما شاهد این بوده ایم که حتی شرکتهای خدمات آماری خارج از کنسرسیوم ما، CRISP-DM را به خدمت گرفته اند ، از این به عنوان یک مدل رایج و استاندارد در صنایع استفاده شده واهمیت آن در میان مشتریان بارها یادآوری شده است. (CRISP-DM هم اکنون در انجام مزایده ها و مناقصه های تجاری نقش تعیین کننده ای را بازی می کند.) ما معتقد هستیم، ابتکار ما در این زمینه کاملا مورد تایید مصرف کنندگان بوده و نیز بسط و توسعه روش هایمان در بروز رسانی این برنامه ناگزیر و لازم الاجرا است. از لحاظ ما، CRISP-DM نسخه 1.0 کاملا بروز رسانی شده و آماده انتشار و توزیع بین مصرف کنندگان است.

CRISP-DM از لحاظ اصول فنی و کاربردی، فقط به مباحث تئوری و آکادمیک نپرداخته است و نه اینکه ادعا کنیم که توسط گروهی فرهیخته از استادان فن در پشت درهای بسته خلق شده است. تجربه ثابت نموده است که این دو راهکار در توسعه روش ها، در گذشته مورد آزمایش قرار گرفته اند و کمتر به نتیجه ای مطلوب (یعنی تولید استانداردهای موفق و قابل قبول و مورد استناد) منجر شده اند. موفقیت CRISP-DM را تنها باید مدیون تکیه گاه محکم عملی و تجربی آن در جهان واقعی دانست و نیز کسانیکه از آن در پروژه های داده کاوی بهره برده اند. در این رهگذر، ما خود را بی نهایت مدیون کارورزان و کاربران انبوهی می دانیم که با تشریک مساعی وابلاغ ایده ها و نظرات خود ما را برای به پایان رساندن این پروژه یاری نمودند.

کنسرسیوم CRISP-DM

CRISP-DM 1.0

پیت چپمن (شورای ملی مقاومت)، جولیان کلینتون (SPSS)، رندی کربر (شورای ملی مقاومت)، توماس خابازا (SPSS)، توماس رینارتز (دایملر کرایسلر)، کالین شیرر (SPSS) و رودیگر ویرث (دایملر کرایسلر)

هدف از این مدرک ، توصیف و تشریح مدل فرآیند CRISP-DM به عنوان یک ابزار داده کاوی می باشد که شامل، مقدمه ای بر شناخت سبک ها و روش های آن ، معرفی CRISP-DM به عنوان یک الگوی مرجع، راهنمای کاربران و ارائه نمونه گزارشات گرفته شده با استفاده از این ابزار نرم افزاری است. در پایان مطالب نیز فهرست و اطلاعات سودمند و مرتبط با موضوع گنجانده شده است.

اطلاعات و مدارک ارائه شده، در مالکیت انحصاری شرکای کنسرسیوم CRISP-DM است که اعضای آن عبارتند از:

NCR Systems Engineering Copenhagen (USA and Denmark)،

DaimlerChrysler AG (Germany)، SPSS Inc. (USA) and

OHRA Verzekeringen en Bank Groep B.V (The Netherlands)