7C.4 ARMY RDT&E METEOROLOGICAL ARCHITECTURE FOR DATA ARCHIVAL (ARMADA)

Scott F. Halvorson*, Susan E. Krippner, Carissa L. Klemmer, Edward P. Argenta,
Margaret B. Kimball and Eric J. Laufenberg
U.S. Army West Desert Test Center, Dugway UT

1. INTRODUCTION

The Army Test and Evaluation Command (ATEC) primary mission is to test and evaluate Army materiel, which may include, but not all inclusive, chemical and biological detectors, rockets, artillery and mortars, explosives, dispersion of smoke and obscurants, electronics and communications. Most of these tests are conducted outside where the environment greatly influences the results. A command group within ATEC called the Army Research Development Test and Evaluation (RDT&E) Meteorology Program supports these tests by providing weather forecasts, forensic analysis, instrumentation, and climatologies at seven Army test facilities across the United States. In particular they collect, archive, and distribute meso and microscale meteorological data at each range. Data sets include measurements from spatially dense networks of surface weather stations, field meters, sonic anemometers, sodars, wind profilers, rawinsondes, and other instrumentation. Timely management and integration of these data are critical for real time modeling (initialization of mesoscale numerical weather prediction models, sound propagation for artillery, mortar, ground vehicle trafficability and explosive tests, and dispersion of military smoke and obscurants, etc.). Additionally, climatological studies require long historical records of these data. The Army RDT&E Meteorological Architecture for Data Archival (ARMADA) was developed to meet the demands for increasing data volume, integration, near real time distribution, and automated quality control

2. DATA COLLECTION

Table 1, shows a synopsis of meteorological observing stations at Dugway Proving Ground (DPG) West Desert Test Center (WDTC). Historically, data collection of these systems has generally been developed around the observing unit, which is driven by mission requirements. These collection platforms typically consisted of a small database or text files on individual computer systems, and often required custom software to translate the data from the field to a usable format. Furthermore, the units are typically a mixture of English and metric, and time can be archived in local standard, local daylight savings, or Universal Coordinated Time (UTC).

The rational for these collection mechanisms started when one of the first deployed atmospheric digital measuring system, Surface Atmospheric Measuring Systems (SAMS), came online in the late 1980's, when computer processing, storage, and data communication resources were limited. Thus, the building of data collection systems around other platforms continued over the last couple of decades, which followed a paradigm of developing the data collection system around the observing platform. Since that time, computer processing, data storage, and networking resources have increased exponentially, but because the data requirements did not change, this paradigm did not change either. Recently, there has been an increasing need to integrate data into numerical models, eliminate confusing labels and units, and reduce labor costs of data management, all of which requires a new paradigm of centralizing all data into one location and enforcing standardization of all of data formats, labels, and units. This new paradigm led to the development of ARMADA, as illustrated in Figure 1.

3. ARMADA

3.1 Data Warehouse

At the core of ARMADA is a data warehouse as defined by Witten and Frank (2005), "Data warehouses provide a single consistent point of access to corporate organization data." ARMADA includes all range meteorological data and associated metadata which is organized by a relational database management system (RDBMS). This provides two functions; 1) all formerly isolated databases, text files, and other sources now centrally located and controlled, and 2) all data can be related to its metadata. Additionally, the database

*Corresponding author address: Scott F. Halvorson, West Desert Test Center, U.S. Army Dugway Proving Ground, Dugway UT 84022. email: scott.halvorson@us.army.mil

Platform Observing Unit	Spatial		Frequency (Seconds)	Observing Parameters (Count)	Observation Records Per Day
	Horizontal (Units)	Vertical (Levels)			
Ceilometers	3	20	15	1	345,600
Electric Field Meters	28	1	1	4	2,419,200
SAMS (Mesonet)	26	1	300	30	7,488
Present Weather	8	1	10	8	23,040
Wind Profiler	2	30	1800	3	2,880
PWIDS	113	1	10	7	976,320
SODAR	3	40	900	2	11,520
Towers (32M)	6	5	10	7	259,200
Sonics (3D)	50	1	0.1	4	43,200,000
Rawinsonde	5	>1000	N/A	8	N/A
Tethersonde	2	5	10	6	86,400

Table 1. A sample list of observing platforms at DPG-WDTC that describes the typical spatial, temporal, and measurements per record of observation. The yellow background rows shows continuous measurements, orange rows show measurements during field tests, and the light blue rows are manual assisted observations. The "Spatial/Horizontal" column shows the maximum possible deployable units, the "Spatial/Vertical" column show the maximum possible vertical levels, and the "Frequency" column is the typical sampling rate of the observing unit during field tests.

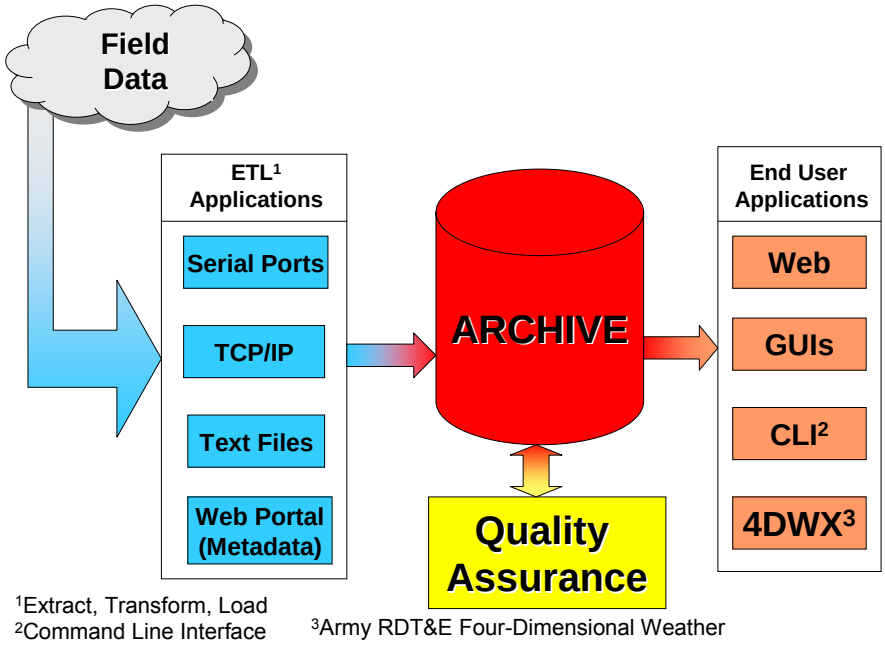

Figure 1. This figure shows the four components of ARMADA; 1) the archive, 2) Extract, Transform, Load (ETL) applications, 3) end user applications, and 4) the quality assurance program. The arrows indicate the flow of data through ARMADA starting from the unprocessed field

structure simplifies the pairing of data and metadata. Historically, metadata has been organized on desktop computer(s). This has resulted in severe data retrieval limitations, and has substantially increased the difficulty in matching data with metadata.

The ARMADA data warehouse is a collection of MySQL AB. databases that are controlled by a single MySQL AB. server which provides input/output access through a single point. A disadvantage is that this architecture can be a single point of failure, but adhering to proper information assurance management procedures defined by Army Regulation 25-2, failures can be greatly reduced.

A key design of ARMADA is that all of the databases and tables are self describing, that follow a predefined naming convention which describes the contents and date validity of the data. For example, the table name "sams_2007" tells the end user it is "SAMS" data, and that it contains all "SAMS" data for the year 2007. The variables (column names) and units follow the Climate Forecast (CF) standard convention, which definitively defines the variable meaning and its associated unit.

3.2 Extract, Transform, Load (ETL)

ETL terminology is often referenced in data warehouse definitions as the process that extracts incoming data, transforms it into an organizations' needs, and loads the data into a target location such as a database. ARMADA applies ETL processes by using stand alone applications that, 1) extract elements from the observation strings in the data stream, 2) transform data to the correct units and calculates derived variables, and 3) loads this data to a MySQL AB. server which populates it into the correct table(s). These applications can be broken down into two types based on the format of the observations, single line or multiline. Single line ETL's are independent of the data format, but designed around different data communication or storage methodologies, such as TCP/IP, serial ports, and text files. A configuration file tells these ETL applications how to interpret the data, convert units, calculate derived variables, and where to store the process data. Multiline observations, such as rawinsonde and SODARS have proven to be more difficult to design into a single application, and therefore have been built around multiline datasets. In the future, both single line and multiline applications may be merged into a single application, or into an application programmers interface (API).

A core concept of ARMADA is that all archived data units must adhere to CF units, which closely follow the International System of Units (SI). It is not always feasible to convert all units at the data logger to CF units, and therefore, a callable library must be utilized by ETL applications to convert these units during the "Transform" phase. A library was designed and implemented to convert units to CF, and then back to non CF units. For example, the CF unit for Temperature is Kelvin, thus the library can convert Celsius, Fahrenheit, and Rankine to Kelvin, and vice versa. Another library component was implemented to compute derived variables during transformation, which is especially useful on more complex calculations that require elements outside of the observation record (i.e., mean sea level pressure).

The relationship between data and metadata must not become out of sync, therefore it is imperative that the metadata be properly maintained in the database. The Portable Weather Instrumentation Data System (PWIDS) is a good example of the necessity to maintain metadata, because these systems often move weekly to monthly, and sometimes hardware needs to be swapped out. These frequents changes to the metadata are prone to delays in recording of these data, which can give false information if not updated in near real time. To assist the field technicians, a web portal was developed, where they could enter metadata information from any available web client on the network. Future plans include developing a program for a personal digital assistant (PDA) which would allow the technicians and meteorologists to enter this information in near real time.

3.3 End User Applications

The creation of a cohesive database is pointless without providing the end user interfaces to the data or a method of distributing data. Typical user interfaces include Graphical User Interfaces (GUIs), manual queries through a console, and other GUI tools. GUIs can be broken down into two components, 2-D displays and data retrieval tools. Figure 2 is an example of a 2D display that continually interrogates the database for current data. In this instance a real-time analysis of the vertical component of the surface electric field is created by retrieving data from the database and subsequently analyzing it (Kimball 2008). An ap-

plication to input real-time data into the Defense Threat Reduction Agency Hazard Prediction and Assessment Capability (HPAC) Second Order Integrated Puff (SCIPUFF) model is another example of a data retrieval tool. This WDTC developed tool interrogates the ARMADA repository, and integrates the user requested data into an HPAC observation file format for input into the HPAC-SCIPUFF model. In addition to 2-D displays and GUI tools, direct logins through a console or query GUI tool are permitted by privileged users to allow execution of sequential query language (SQL) statements. This type of access is necessary for complex climatology studies or other more specific data retrieval.

Another category of services which interfaces with ARMADA, are noninterface services. These are data pushers that continuously push data to an end user. An example of this type of application is a data pusher that pushes current SAMS data to the Western Region Headquarters of the National Weather Service in Salt Lake City, Utah allowing public viewing and data access via the National Weather Service webpage.

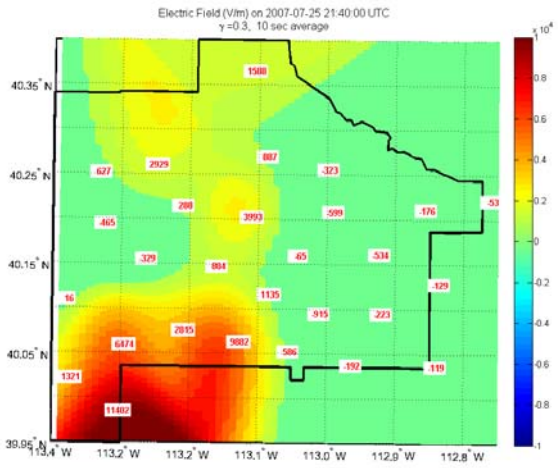

Figure 2. An example of a horizontal analysis of the vertical component of the electric field over DPG with data derived from ARMADA. The values listed in the image indicate the electric field strength (V/m) at each observing station for the time period indicated at the top of the image.

3.4 Quality Assurance (QA)

The goal of ARMADA is to archive all data transmitted by the sensor station. Therefore, data must be quality controlled (filtered or flagged) prior to distribution to the end user. The QA program goal for ARMADA requires automated near real-time and manual QA while maintaining the requirements that all original data must be preserved.

Automated QA is necessary for near real time displays and models. Such a system will reduce data delivery turn-around time and labor costs during field tests. Because automated QA is not perfect, a QA manager must be able to overturn or apply any flagged results. A flow diagram of the mechanics of the QA process that is currently in the development phase at DPG is shown in Figure 3.

QA is only performed after the data is populated into the database because QA would be too taxing on ETL applications. Furthermore, if there was found to be an incorrect algorithm or an additional QA procedure, it would be very difficult to retroactively QA “live” data. Therefore, a separate application called the “Quality Control Server” (QCS), shown in Figure 3, runs independently of ETL. The QCS utilizes a programmable configuration file that determines the frequency, methods, data types, and other configurations for proper QA testing. For example, these tests can check the validity of the data by means of comparing the range of measurement to climatic possibilities, or to look for changes in value between measurements. The end result of the QA tests are flags that describe any tests that failed. These results then get published back into the data table. If the data passes all tests, it also gets written to a “gold standard” table. Most of the time, the end user just needs the best data, and would therefore use the gold standard table. This allows for more efficient data processing.

4. FUTURE

ARMADA is anticipated to become fully operational for this upcoming test season (early spring 2008). However, ARAMADA is expected to rapidly grow over next three to five years to include new datasets, ETL capabilities, end user applications and distribution, and QA. New datasets will likely include Radar, LIDAR, and other instrumentation that produce non-ASCII formatted data structures. Similarly, other data sets may include

non-range data, such as data from NOAAPORT. These new data sensors will require new ETL applications and additional or modification of end user applications. QA will initially be able to run a few tests primarily on surface-based observations, which typically include SAMS and PWIDS. Wind profiler and SODARs are expected to use the NCAR Improved Moments Algorithm (NIMA), and sonic data will continue to be quality controlled by in-house software.

Future plans for ARMADA include an application that could continuously encode data into NetCDF or GEMPAK format with distribution through Unidata's Local Data Manager (LDM) as well as building a web server to prevent flagrant SQL commands from over taxing the database server causing a "Denial of Service".

5. SUMMARY

Due to current mission requirements for data integration and timely management, the paradigm of isolating datasets has shifted to a new a concept of centralizing data which ARMADA accomplishes. In order for ARMADA to implement this approach, it needs the four basic components, 1) centralized archive, 2) a method to upload field data to the repository (ETL), 3) the capability to utilized the data in the central archive, and 4) the ability to QA data.

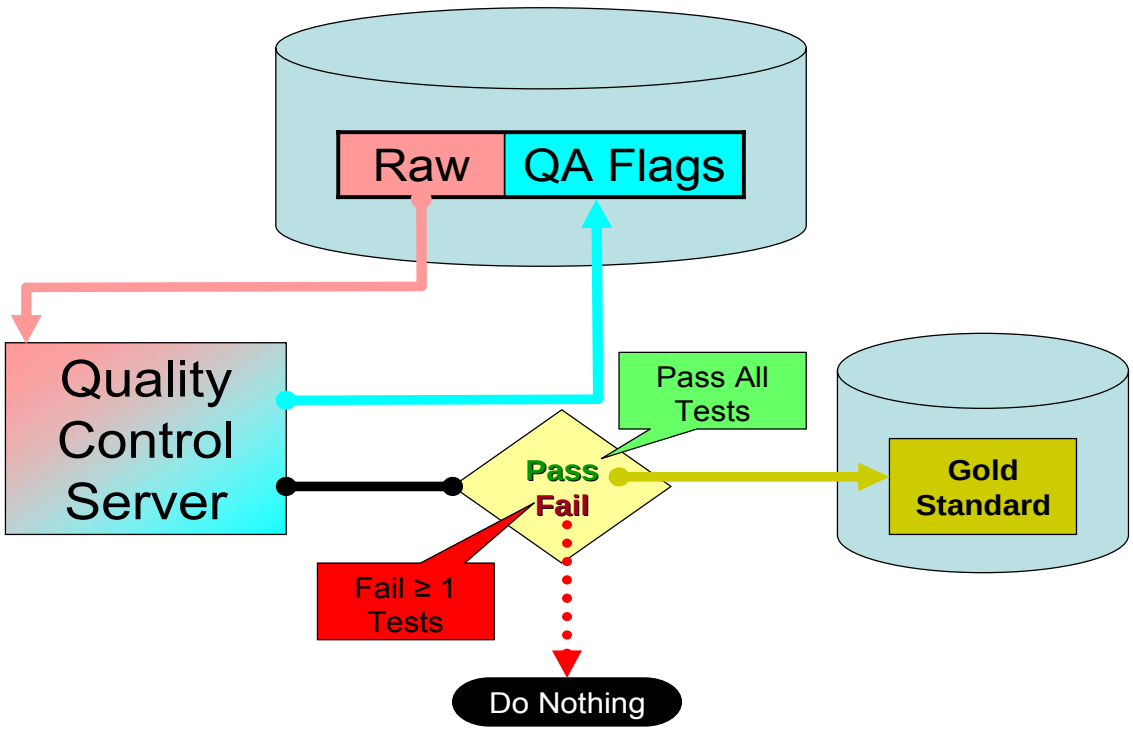

Figure 3. This figure illustrates the basic concept of the ARMADA QA process.

6. REFERENCES

Climate Forecast standard names. Available from http://cf-pcmdi.llnl.gov/documents/cf-standard-names/6/cf-standard-name-table.html

Friebrich, C. A., and K. Crawford, 2001: The impact of unique meteorological phenomenon detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. *Bulletin of the American Meteorological Society*, **82**, 2173-2187.

GEMPAK, LDM, and NetCDF applications and libraries are provided courtesy by Unidata and are available from http://www.unidata.ucar.edu/

Kimball, M. B. and F. W. Gallagher, 2008**:** Operational evaluation of lightning precursors from a network of field mills at Dugway Proving Ground. Preprints, 3rd Conference on Meteorological Applications of Lightning Data, New Orleans LA, Amer. Meteor. Soc., P2.11.

Kimball, R. and J. Caserta, 2004: The data warehouse ETL toolkit. Indianapolis, IN: Wiley Publishing Inc.

MySQL AB. Database Server. Available from http://www.mysql.com

NOAAPORT. Available from http://www.nws.noaa.gov/noaaport/html/noaaport.shtml

Shafer, M. A., C. Fiebrich, D. Arndt, S. Fredrickson, T. Hughes, 2000: Quality assurance procedures in the Oklahoma Mesonet. *Journal of Atmospheric and Oceanic Technology*, **17**, 474-488.

Witten I. H., and E. Frank, 2005: Data mining practical machine learning tools and techniques. Boston: Morgan Kaufman Publishers.

Platform Observing Unit	Spatial		Frequency (Seconds)	Observing Parameters (Count)	Observation Records Per Day
	Horizontal (Units)	Vertical (Levels)			
Ceilometers	3	20	15	1	345,600
Field Mill	28	1	1	4	2,419,200
SAMS[1]	26	1	300	30	7,488
Visibility/Precip	8	1	10	8	23,040
Wind Profiler	2	30	1800	3	2,880
PWIDS[2]	113	1	10	7	976,320
SODAR	3	40	900	2	11,520
Towers	6	5	10	7	259,200
Sonics (3D)	50	1	0.1	4	43,200,000
Rawinsonde	4	>1000	N/A	8	N/A
Tethersonde	2	5	10	6	86,400

Table 1, is sample list of observing units at DPG-WDTC that describes the typical spatial, temporal, and measurements per record of observation. The yellow background rows shows continuous measurements, orange rows show measurements during field tests, and the light blue rows are manual assisted observations. The "Spatial/Horizontal" column shows the maximum possible deployable units, the "Spatial/Vertical" column show the maximum possible vertical levels, and the "Frequency" column is the typical sampling rate of the observing unit during field tests.

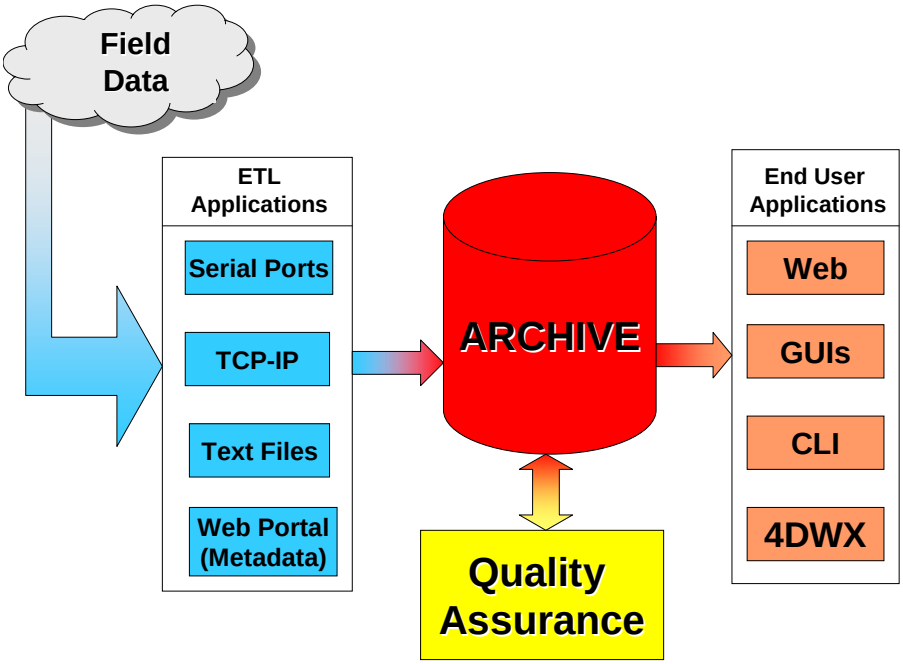

Figure 1, shows the flow of data into and out of ARMADA.

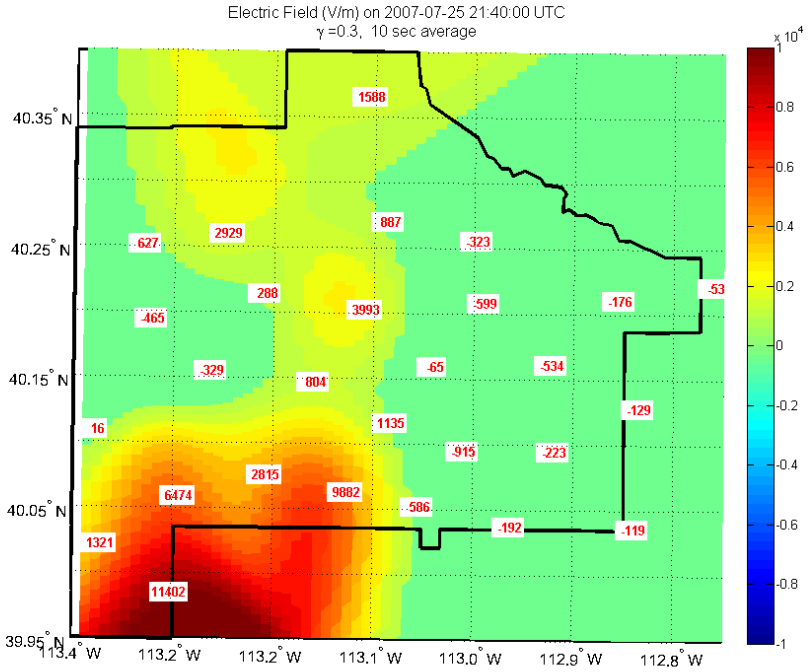

Figure 2, shows an example of a horizontal analysis of the vertical component of the electric field over DPG. The values listed in the image indicate the electric field strength (V/m) at each observing station for a given period defined at the top of the image.

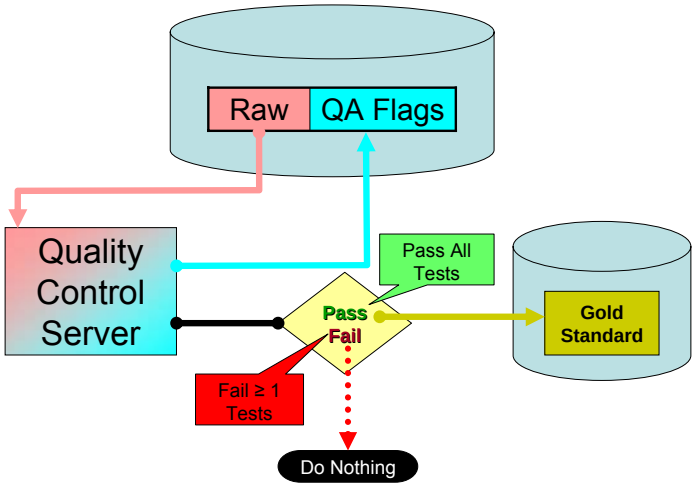

Figure 3, shows the ARMADA QA flow diagram.